DOCUMENT RESUME

ED 431 815 TM 029 935

AUTHOR Loomis, Susan Cooper; Bay, Luz; Yang, Wen-Ling; Hanick, Patricia L.

TITLE Field Trials To Determine Which Rating Method(s) To Use in the 1998 NAEP Achievement Levels-Setting Process for Civics and Writing.

INSTITUTION ACT, Inc., Iowa City, IA.
SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE 1999-04-22
NOTE 63p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999). For related document, see TM 029 934.
CONTRACT ZA07001001
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Academic Achievement; *Academic Standards; *Civics; Elementary Secondary Education; Evaluation Methods; *Field Studies; National Competency Tests; Pilot Projects; Standardized Tests; Tables (Data); *Writing Achievement; *Writing (Composition)
IDENTIFIERS *National Assessment of Educational Progress; Standard Setting

ABSTRACT

Field trials were conducted to test rating methods and the impact of feedback about consequences for the 1998 achievement levels-setting (ALS) process for the National Assessment of Educational Progress (NAEP). The field trials provided the opportunity to try out different methods similar to those used successfully by others, as well as to try out some new methods. The American College Testing program (ACT) had proposed a new method to be tested in the field trials. Although successful implementation of the method had been reported, the method was found to be biased, and the ACT stopped tests with the method after the first field trials. Reservations about item maps were not overcome in the field trial process, and item maps were eliminated as a choice. Concerns about computational procedures and the logistic demands of the Booklet Classification Method eliminated this approach. The Technical Advisory Committee on Standard Setting recommended a new combination method based on the method developed by M. Reckase in conjunction with the strong research base and extensive experience by ACT associated with the Mean Estimation method. Procedures based on this approach were used to set achievement levels for the 1998 NAEP in civics and writing. Appendixes contain examples of charts used in the rating method study. (Contains 18 tables, 33 figures, and 25 references.) (Author/SLD)

Field Trials to Determine Which Rating Method(s) to Use in the 1998 NAEP Achievement Levels-Setting Process for Civics and Writing

by

Susan Cooper Loomis
ACT, Inc.

and

Luz Bay
Advanced Systems

Wen-Ling Yang
ACT, Inc.

Patricia L. Hanick
ACT, Inc.

Presented at the Annual Meeting of the
National Council on Measurement in Education

Montreal
April 22, 1999

Field Trials to Determine Which Rating Method(s) to Use in the 1998 NAEP Achievement Levels-Setting Process for Civics and Writing[1]

Susan Cooper Loomis, ACT, Inc.
and
Luz Bay, Advanced Systems
Wen-Ling Yang, ACT, Inc.
Patricia L. Hanick, ACT, Inc.

Introduction

ACT proposed several stages in preparation for the 1998 Achievement Levels-Setting (ALS) Process (ACT, 1997). These included simulation studies followed by field trials, followed by pilot studies. The major focus of these research studies was identification of a rating method for setting the achievement levels. Because achievement levels-setting is a judgmental process, the best way to improve the outcomes of the process is to improve the method of collecting judgments. In designing the methodology for setting achievement levels for the 1998 Writing and Civics NAEP, ACT did not seek to find a "true standard," rather, ACT tried to design a data collection procedure based on a judgment task that panelists could easily comprehend so that systematic bias in judgments would be minimized.

In an effort to improve procedures used for the 1998 NAEP achievement levels-setting (ALS) process, ACT proposed to use an item-by-item rating method that was somewhat different from the modified-Angoff rating method used in recent years. The change of method was proposed in response to criticisms that the modified-Angoff method could not produce valid cutpoints because panelists were incapable of performing the task of estimating probabilities with reasonable accuracy (NAE, 1993; Shepard, 1995; Impara and Plake, 1996).

ACT proposed a rating procedure that required judges to estimate the most likely response of borderline student performance at each achievement level (ACT, 1997f). This method differed from the modified-Angoff method in that panelists were asked to estimate the most likely response, rather than the probabilities of correct responses of student performance at the borderline of each achievement level. Angoff (1971) described this procedure; Impara and Plake (1997) worked with this rating method with dichotomous items; and Hambleton and Plake (1995) used a procedure somewhat similar in a standard setting study involving polytomous items. These studies reported success with using the method.

ACT conducted the simulation studies (Chen, 1998) and determined that the proposed method (and computational procedures developed for it) was feasible in the context of NAEP. ACT called the method the "Item Score String Estimation (ISSE) Method because the ratings would produce an "item score string" for students performing at the borderline of each achievement level. For dichotomous items, panelists would judge whether students performing just at the borderline at the achievement level were more likely to respond correctly or incorrectly. For polytomous items,

1 This research was conducted under contract ZA07001001 with the National Assessment Governing Board. Susan Loomis wrote this report, but the report draws heavily upon earlier reports by the author, Luz Bay and Patricia Hanick. Wen-Hung (Lee) Chen at ACT developed the analyses programs for these field trials and helped with on-site analyses. Wen-Ling Yang performed the analyses and produced feedback for the studies, as well as additional analyses for reporting on the studies. Teri Fisher at ACT coordinated the acquisition and production of materials for each of the studies and assisted with materials in this and earlier reports. Jill Crouse at ACT conducted the analysis of FT2 data and prepared summary reports to share with our Technical Advisory Committee on Standard Setting.

they would judge the most likely score (e.g., 1-4) for students performing just at the borderline of each level.

The computations required to determine the cutpoints were simplified using the proposed ISSE method. ACT's proposed new method combined two forms of assessment items; judges would provide expected scores for performance items, and correct/incorrect scores for multiple-choice items. Previously, cutpoints needed to be computed separately for dichotomous and polytomous items. Concerns had been raised regarding how to combine ratings for items that were generated from different rating methods. The ISSE method eliminated this concern.

Overview of the Field Trials

Two field trials were originally proposed in which research studies were conducted with panelists to determine which rating method to use in the ALS process. Each field trial had a unique purpose, but they both address the common issue of examining rating methods. NAGB asked ACT to plan separate field trials for writing and for civics and to examine more alternative methods for setting achievement levels in writing. As a result, ACT scheduled a pair of field trials for each of the two subjects.

This was the first time that field trials, i.e., studies with panelists, had been conducted prior to the pilot studies. In 1994, ACT conducted the pilot studies for geography and U.S. History with major research components included. Four different rating methods were tried out during those two pilot studies, and there were variations in feedback provided to panelists. The decision of which rating method to use needed to be made prior to the pilot studies, and the decision needed to be informed by research involving panelists. ACT and TACSS felt that it was important to conduct the research for identifying methods *prior to* the pilot studies so the pilot studies could be "dress rehearsals" for the ALS process. Thus, the field trials were included in the 1998 process. As happens so often, the field trials grew in scope and complexity as the details of the designs were being worked out.

The initial purpose of the first field trial (FT1) was to compare the ISSE method to the combined methods of modified-Angoff and mean estimation (ME), which had been used by ACT in ALS processes for geography, U.S. History, and science. (See ACT, 1997f.) Results from FT1 were to determine which item-by-item rating method to use for the remaining ALS studies, including the second field trial (FT2). The selected method was to be used in civics and perhaps writing. By the time the field trials were actually designed, however, four methods were being considered for FT1 for writing: the mean estimation method, the ISSE, the Booklet Classification Method, and a new method named The Grid Method.

As proposed, the key issue for FT2 was to study the ratings produced from panelists using a sequence of rating methods: one method followed by a second, different method. The first method would be the item-by-item method selected as a result of FT1 (either ISSE or the combined methods of modified-Angoff and ME), and the second method was to be an item-mapping method. Field trial 2 was to address several issues: how the two methods interfaced with each other when used together in a rating sequence; how the panelists evaluated the two methods when used jointly and independently; and how the cutscores that were produced by the two methods differed.

The effect of consequences data on panelists' ratings was also investigated in both FT1 and FT2. NAGB had never approved the introduction of consequences data during the rating process—before the final cutscores were computed. ACT began providing panelists with consequences data in 1994, based on the final cutpoints, and collecting panelists' reactions to the data. Those data were reported to NAGB and considered during their deliberations regarding the cutpoints to set

for each achievement level in geography, U.S. History, and science. Indeed, it was the reaction of grade 8 science panelists to the consequences data that led to the decision to reconvene the panel and provide them the opportunity to reconsider their cutpoints (ACT, 1997c).

The field trials were designed to collect data to determine the extent of impact by consequences data on ratings. ACT needed to know when to introduce consequences data and how often to provide the data if they were to be used in the process.

Field Trial 1: A Comparison of Two Item-by-Item Rating Methods

The purpose of the field trials was to determine the rating method to be used to set achievement levels for the 1998 Civics NAEP and the 1998 Writing NAEP. The Technical Advisory Committee on Standard Setting (TACSS) recommended that a minimum of ten panelists be recruited for each method group. Because ACT was not able to recruit that many panelists for the scheduled dates of the field trials, the design of FT1 in each subject was modified. Only the ISSE method was implemented for the FT1 in civics and only the ISSE and ME methods (as originally planned) were implemented for FT1 in writing. Consequences data were introduced before the final cutpoints were set for both subjects in FT1.

Data

ACT used items from the 1994 Geography NAEP in the field trials for civics and 1992 NAEP Writing data for the field trials for writing. ACT wanted to include feedback to panelists in the field trials, so it was necessary to use NAEP data that were already available. The 1998 assessment data were still being collected at the time of the field trials.

The Geography NAEP was quite similar to the Civics NAEP in terms of the types of items (multiple-choice, short constructed response, and extended constructed response) and the relative frequency of each. Further, neither geography nor civics represents a "core" course in the curriculum, and the two were judged to be similarly represented in the curricular offerings of schools at the grade levels tested by NAEP. Of all of the subjects for which achievement level data were available, geography seemed the most logical substitution for civics. ACT used the achievement levels descriptions for the 1994 Geography NAEP in the field trials. Panelists were asked to avoid reference to any reports on the Geography NAEP prior to participation in the civics field trials.

ACT had worked with the 1992 Writing NAEP data, and that seemed the most obvious choice of data to use for the field trials. The problems experienced with the assessment data in 1992 were of concern, however. The framework document had been revised somewhat, and the test specifications had been sharpened and tightened since the 1992 assessment. ACT used the achievement levels descriptions that had been developed for the 1998 Writing NAEP in the field trials for writing. The generic scoring rubrics for 1998 were specifically worded to avoid using exactly the same terms used in the achievement levels descriptions. The correspondence, or lack thereof, between the 1992 scoring rubrics—specific to each prompt—and the 1998 ALDs was not taken into account for the field trials.

Rather than use all of the items in the 1994 geography assessment, ACT used only the four blocks of items that had been used in validation studies for the geography achievement levels. Those studies were conducted with data for grade eight only (ACT, 1995). The four item blocks were selected to maximize the representation of the content framework and the characteristics of the entire item pool for grade eight (Carlson, 1995). This choice to use only four representative blocks was made to decrease the amount of time required by the rating process.

There were no similar concerns about time for rating writing prompts. The 1992 Writing NAEP for grade 8 included only 11 prompts in all, and only 9 of those were 25-minute prompts. Since only 25-minute prompts were to be used in reporting the 1998 NAEP, only the nine 25-minute prompts from the 1992 NAEP for grade 8 were used in the field trial.

Panelists

Twenty persons were to be empanelled in Iowa City for FT1 in civics and 40 persons were to be empanelled for writing: 10 panelists for each rating method in the trials. The process for each subject was planned to last two days. Panelists were recruited from the counties in eastern Iowa, around ACT's national headquarters. The recruiting process was somewhat similar to that planned for the actual ALS and pilot studies in that persons in specific positions (superintendents, curriculum supervisors, mayors, and so forth) were asked to nominate teachers, nonteacher educators, and general public representatives to serve on the panels. Panels were to be drawn from the nominees to optimize the composition with respect to the targeted demographic attributes for panels. Our highest priority is given to selecting panelists with the best qualifications. NAGB specifies that the panels are to include three types of judges, and 55% of the panelists should be teachers, 15% nonteacher educators, and 30% general public representatives. (Please see 1997g for details.) Panelists were offered a small honorarium of $100 to participate in field trial #1.

The first field trial for civics was conducted February 7-8, 1998 and the first field trial for writing was conducted February 28-March 1, 1998. ACT contacted hundreds of persons and asked for nominations: school officials, elected officials, and companies likely to employ persons actively engaged in working with knowledge and skills related to the subject areas. Despite intensive efforts to recruit panelists, only 8 persons could be recruited for the first field trial for civics and 15 for writing. Teachers were simply too busy during this part of the school term to participate in these studies. Further, the curriculum supervisors suggested that teachers are unwilling to spend two weekend days working for such a small fee.[2]

The civics panel included 2 current teachers and one in her first year of retirement, 3 nonteacher educators, and 2 general public members. There were two men and five women. The composition of the writing FT1 panel was unique. For the first time ever, more general public members than educators were included on a NAEP ALS panel. The writing FT1 panel included 4 teachers, 3 nonteacher educators, and 8 members of the general public. There were six men and 9 women in the writing FT1.

Process

Training. The NAEP ALS process typically lasts five days. All aspects of the ALS process, except selection of exemplar items, were covered in the two-day field trials, at least to some extent. ACT wanted to ascertain how panelists reacted to the rating methods and to other procedural changes proposed for the 1998 NAEP ALS process, so some aspects of the process were sacrificed in the context of collecting the field trial data. Relative to the typical ALS process, field trial time was greatly reduced for training in the frameworks and achievement levels descriptions. Further, there were only two rounds of item ratings.

Panelists were provided an abbreviated orientation to the achievement levels-setting process and the process designed for the field trial. The orientation session included a general orientation to the NAEP program and the process of developing NAEP achievement levels. Panelists participated in several training exercises that were the same as those provided to ALS panelists.

2 ALS panelists are paid no honorarium, and the $100 had been judged "appropriate" for field trial panelists. Field trial #2 panelists were paid $300 for the two-day study.

Included in the training was administration of a form of the NAEP for each panelist to complete. By taking the NAEP and scoring their work, participants become familiar with specific NAEP items and scoring rubrics and with the general format of the assessment and the conditions under which it is administered.

Since only 8 panelists were recruited for civics, only the ISSE method was implemented. ACT judged, and the Technical Advisory Committee on Standard Setting (TACSS) concurred, that the results of the ISSE method could be compared to the data collected in the geography ALS process using the Mean Estimation Method. ACT already had considerable experience with the ME method in assessments similar to civics. Both the Mean Estimation and the ISSE were implemented in the FT1 for writing, but the other alternatives being considered had to be eliminated for the first writing field trial.

Panelists spent the first day in training and preparation for rating items at the end of the first day. Panelists reviewed assessment items, scoring rubrics, and student papers, and they were engaged in exercises to become more familiar with the achievement levels descriptions before the first rating session. The process implemented for each subject was quite similar, but some adjustments were needed to accommodate specific features of the assessments in the two different subjects.

ACT typically uses a paper selection exercise to train panelists in the scoring rubrics for polytomous items, to give them a "reality check" prior to the first round of ratings, and to give them experience in applying their concept of borderline performance with respect to student performance. The paper selection process was not implemented in the civics field trial, but it was implemented in the writing field trial with papers written in response to three prompts, one of each of the three types of writing assessed by NAEP. Three student papers were included for each of 6 score points for a narrative prompt and for an informative prompt; the two highest score points had been collapsed for the persuasive prompt and only papers for the 5 score points were included. Each panelist thus had 51 student papers from which to select one paper to represent borderline performance for each achievement level.

Ratings and Feedback. There were two rounds of item-by-item ratings. Panelists were asked to form a concept of students performing at the borderline of each level. For the ISSE method, they were asked to judge whether such students were more likely to answer each multiple-choice item correctly or incorrectly. For constructed response items, they were asked to estimate the most likely score for such students.

Panelists reported no special problems with the rating methodology, and the first round of item ratings went smoothly for civics. Writing panelists had more trouble with the rating methods, and this seemed especially true for some panelists in the ME group. Writing panelists had difficulty reconciling the scoring rubrics with the scores they had seen for some student papers in the paper selection process. They also had problems with the scoring rubrics relative to the achievement levels descriptions and the limited amount of time (25 minutes) allowed for student responses.

The first round of ratings was collected in the afternoon of the first day. After rating all items in their rating pools, panelists left for the day. The rating forms were collected for computation of cutpoints and other feedback information overnight. All cutpoints reported to panelists and presented here are reported on the ACT NAEP-Like scale.

Prior to the second round of rating on the second day of FT1, panelists were given feedback data resulting from the ratings they provided in the first round. For FT1 writing, separate sets of cutpoints and other feedback were computed for panelists in each of the two methods groups.

Panelists were provided with cutpoints, rater location charts, p-value tables, and whole booklet feedback. They were instructed in the source, meaning, and use of the feedback information. These feedback data were the same as used in previous NAEP ALS processes (ACT, 1997a, 1997b, 1997c). Rater location data are provided as charts showing the location of the cutscore for each panelist at each achievement level. P-value tables report the percentage of students answering each dichotomous item correctly and both the average score for polytomous items and the percentage of students scoring at each rubric point. The whole booklet feedback reports the expected percent correct score for the set of items in the NAEP exam booklet that panelists took earlier for practice. For example, the whole booklet feedback report might state: "Based on your group's average ratings, students performing at the borderline Basic level are expected to get 49% of the total possible score points for this booklet." (A similar statement is given for each achievement level.) This feedback was based on the cutpoints the group had set during the previous round of ratings.

Panelists also participated in the whole booklet exercise, an extension of the whole booklet feedback. This exercise was added to the ALS process in 1994 in response to the NAE (1993) recommendation to include more "holistic" procedures in the process. To illustrate borderline Basic performance, they were shown copies of booklets with scores around 49% of the total possible points, for example. A few booklets scored within 2% of the cutpoint of each achievement level (above or below) were shared with panelists for their evaluation. They were asked to examine the responses of students and determine whether that performance represented their expectations for students at the lower borderline Basic level, for example. If they perceived a discrepancy between the performance expected and observed in the booklets scored at the cutpoint, then they were to discuss the achievement levels descriptions and borderline performances again with other panelists and try to understand the cause for this discrepancy. They were told that if they judged the performance to be too low relative to the description for achievement at the level, they should increase their ratings for the level(s). If they judged the performance to be higher than they would expect relative to the description for achievement at the level, then they should decrease their ratings for the level(s).

During the second round of ratings, panelists again rated all items in their item pool. They were told that they could change ratings for any items at any levels. Ratings were collected and feedback data produced for their review within about two hours.

Consequences Data. The feedback information described above was updated after the second round of ratings. The percentages of students scoring at or above each achievement level based on the cutpoints that they set on the second round were provided as consequences data. ACT had proposed to introduce consequences data before the final round of ratings, and this field trial was one of several opportunities planned for collecting data to study the effect of providing consequences data. ACT had collected panelists' reactions to consequences data provided *at the end* of the ALS process, and those data suggested that few panelists would make changes. Still, there was no evidence regarding what panelists would do when their actions could impact the cutscores. (Please see Appendix 1 for an example of the consequences feedback.)

Panelists were asked to complete a questionnaire in which they were given the opportunity to recommend new cutpoints that would raise or lower the percentages of students performing at or above each level. Those numbers were averaged and new cutpoints and consequences data were

presented. Panelists were allowed to discuss the cutpoints and encouraged to reach common agreement on a final set of cutpoints.

The civics panelists had a rather lengthy discussion of the results and consequences, but there was relatively little interest in changing the cutpoints. Four of the eight panelists recommended that the cutpoints be reported as set. One panelist recommended a change in only the Basic cutpoint—lowering it. Two panelists recommended changes in two cutpoints. One of those panelists would lower the Proficient and Advanced cutpoints and one would lower the Basic and Proficient cutpoints. One panelist recommended changes to all three cutpoints.

The writing panelists, on the other hand, seemed generally appalled by the consequences data. Many spoke about their reluctance to "arbitrarily" change the cutscores, although most seemed to find the outcomes unreasonable. When asked whether the results reflected their expectations, only one panelist in the ISSE rating group said "yes." There were many fewer changes recommended by panelists in the ME group for the writing FT1 in response to the consequences data. Five panelists in each group said they would change one or more cutpoints, and four of those five in the ISSE group changed all 3 cutpoints. All changes to the Proficient and Advanced cutpoints by ISSE panelists were to lower the cutpoints and increase the percentage of students scoring at or above the levels. Only one of the four changes to the Basic cutpoint in the ISSE group was to make it higher. In the ME group, only one panelist changed the Basic and Proficient cutpoints, and five panelists lowered the Advanced cutpoint.

The recommended changes were used to compute new cutpoints and revised consequences data. Those were again shared with panelists, and they were again asked to evaluate the data. Members of the civics FT1 panel had no further changes to recommend, but the writing panels did.

When asked whether the revised, "final" percentages reflected their expectations, all panelists in the ME group now said "no," and five in the ISSE group said "no." In the ME group, only one panelist would raise the Basic cutscore and the rest would leave it as set. Five ME panelists would lower the Proficient cutpoint and three would leave it as set. Seven would lower the Advanced cutpoint and one would leave it unchanged. They expressed a general lack of confidence in the "arbitrariness" of the cutpoints computed on the basis of recommendations. There was a general preference for the cutpoints based on their ratings.

Three ISSE panelists would lower the Basic cutpoint and four would leave it unchanged. Two would raise the Proficient cutpoint, one would lower it, and four would not change it.

Results

TACSS reviewed all data from the field trials in both civics and writing. Replicability of results is one criterion TACSS suggested ACT use in evaluating the outcomes of the field trials for selecting a method. In general, the number of panelists was small for placing heavy emphasis on the numerical outcomes. The intent of the studies had been to ascertain how well panelists react to and interact with the methods and the feedback provided for each method. The evaluation data collected from panelists, along with observations of staff engaged in implementing the process were of greatest interest to TACSS. ACT conducted extensive analyses of the data, and only highlights are presented here.

Interjudge consistency evidence and intrajudge consistency evidence were available, to a somewhat limited extent. Plake (1995) suggested that high interjudge consistency (low variability among panelists' ratings) could be used as an indicator of replicability. Table 1 reports the cutscores and standard deviations on the ACT NAEP-Like scale for the civics FT1 ratings and

Table 2 reports the data for both rating methods for the writing FT1 ratings. There were no data to which standard deviations from the civics FT1 could be compared. Ratings for the four blocks in the geography grade 8 pool are reported below, but those data could not be used in computing the standard deviations because of differences in rating groups and rating pools in the ALS process.

Table 1
Cutpoints and Standard Deviations for ISSE Ratings of
4 Blocks of Items in the Grade 8 NAEP Geography Item Pool

	Basic Cutpoints (SD)	Proficient Cutpoints (SD)	Advanced Cutpoints (SD)
ISSE Method (Round 1)	149.62* (7.93)	171.47 (6.73)	189.75 (8.66)
ISSE Method (Round 2)	152.19 (9.79)	171.47 (4.67)	187.33 (4.00)

*Cutpoints are reported on the ACT NAEP-Like score scale.

The data in Table 2 for writing show the standard deviations to be lower for the ISSE than for the ME in writing.

Table 2
Cutpoints and Standard Deviations for ISSE and ME Ratings
of 25-Minute Prompts in the 1992 Grade 8 NAEP Writing Pool

	Basic Cutpoints (SD)	Proficient Cutpoints (SD)	Advanced Cutpoints (SD)
Writing FT1 ISSE Method Round 1	134.87* (7.02)	177.81 (10.18)	229.12 (6.92)
Writing FT1 ISSE Method (Round 2)	137.15 (1.82)	174.24 (5.9)	221.92 (10.52)
Writing FT1 ME Method (Round 1)	147.31 (14.78)	184.15 (12.17)	220.83 (12.05)
Writing FT1 ME Method (Round 2)	142.63 (12.95)	176.39 (10.92)	213.05 (9.47)

* Cutpoints are reported on the ACT NAEP-Like score scale.

Ratings by civics FT1 panelists using the ISSE resulted in higher cutpoints than those computed from ratings by geography ALS panelists on the same items.[3] For writing, the ISSE method produced cutpoints that were more extreme: the Basic cutpoint was lower and the Advanced higher than for the ME method. Since the NAEP ALS cutpoints have generally been criticized as being too high, a method that set even higher cutpoints would not likely be selected, *other things being equal.* The percentages based on the Round 2 item ratings by FT1 civics panelists and Round 3 item ratings by the ALS panelists are reported in Table 3.

Table 3
Percentages of Students Scoring At or Above Cutpoints Set for 4 Blocks of Items in the Grade 8 NAEP Geography Item Pool

	% At or Above Basic	% At or Above Proficient	% At or Above Advanced
Civics FT1 ISSE Method (Round 2)	61.6%	11.6%	0.3%
Geography ALS ME Method (Round 3)	66.3	26.6	5.5

These results *suggested* that the percentages of students scoring at or above the levels would be lower using the ISSE method than the percentages at or above the cutpoints set in 1994 using the ME method. The 1994 grade 8 Geography cutscores based on all items in the grade pool resulted in 71% of the students scoring at or above the Basic level, 28% at or above the Proficient level, and 4% at or above the Advanced level.

The results for the two methods used in the writing FT1 were quite similar, but the differences showed the cutpoints using the ISSE method were slightly lower for Basic and considerably higher for Advanced than those using the ME method. This is, of course, contrary to the indications from the results of FT1 in civics. Both the ISSE and ME cutpoints for writing FT1 were higher than those for grade 8 in 1992 using the paper selection method. The round 2 FT1 percentages of students scoring at or above the cutpoints are reported in Table 4 along with the data from the 1992 ALS process. Please note that the process through which the 1992 ALS results were reached, along with the computational procedures, were different from those used in FT1 for writing.

Another measure evaluated by ACT and TACSS was changes in ratings. Reviewers of the NAEP ALS process often comment that there is little or no change in cutscores from round to round. This observation is often followed by questions regarding the necessity or utility of having 3 rounds of ratings. ACT has found that panelists typically change relatively large numbers of item ratings from round 1 to round 2, and that they change many fewer from round 2 to round 3. A summary of changes by levels is provided for panelists' ratings in each of the two subjects. Table 5 reports the percentages of changes from Round 1 to Round 2 averaged over the 8 panelists for civics and Table 6 reports the data for changes in ratings for writing, by method.

Data in Table 5 show that item ratings at the Basic and Proficient levels were more frequently raised than lowered, but the proportion of ratings lowered at the Advanced level was just slightly greater than that raised. In general, however, more ratings were unchanged from round to round by the civics FT1 panelists.

3 These data are not entirely comparable because panelists did not participate in exactly the same process. Further, the items included in these four blocks were not all rated by the same panelists in the ALS process. The data are presented as a point of comparison.

Table 4
Percentages of Students Scoring At or Above Cutpoints Set for Prompts
in the 1992 Grade 8 NAEP Writing Item Pool

	% At or Above Basic	% At or Above Proficient	% at or Above Advanced
Writing FT1 ISSE Method (Round 2)	89.9%	8.7%	0.0%
Writing FT1 ME Method (Round 2)	82.0	6.2	0.0
Writing ALS Paper Selection Method (Round 3)	85.0	15.4	0.1

Table 5
Percentages of Civics FT1 ISSE Item Ratings Changed
from Round 1 to Round 2
for 4 Blocks of Items in the Grade 8 NAEP Geography Item Pool

% Basic Ratings Raised	% Basic Ratings Same	% Basic Ratings Lowered	% Proficient Ratings Raised	% Proficient Ratings Same	% Proficient Ratings Lowered	% Advanced Ratings Raised	% Advanced Ratings Same	% Advanced Ratings Lowered
15.0%	73.6%	11.4%	13.0%	72.3%	9.2%	4.3%	91.0%	4.7%

Table 6
Percentages of Writing FT1 Item Ratings Changed (by Method) from Round 1 to Round 2
for 8 Prompts in the Grade 8 NAEP Writing Pool

	% Basic Ratings Raised	% Basic Ratings Same	% Basic Ratings Lowered	% Proficient Ratings Raised	% Proficient Ratings Same	% Proficient Ratings Lowered	% Advanced Ratings Raised	% Advanced Ratings Same	% Advanced Ratings Lowered
Mean Estimation (n=8)	7.9%	50.8%	41.3%	0.0%	44.4%	55.6%	3.2%	42.9%	54.0%
ISSE (n=7)	19.4%	68.1%	12.5%	9.7%	69.5%	20.8%	1.4%	77.8%	20.8%

Relative to civics, a much larger *proportion* of items was changed in the writing field trial. The majority of item ratings were unchanged for panelists using the ISSE method. Panelists who used the ME method changed a larger proportion of items than panelists who used the ISSE method.

Panelists using both methods tended to lower their ratings more frequently than to raise them, and this was especially true for Proficient and Advanced level ratings.

Evaluations of Panelists. To better understand panelists' perceptions of the rating process, they were asked to respond to three questionnaires with Likert-type scale items—one questionnaire after each round of ratings and another at the conclusion of the meeting.[4] Evaluation data from both the geography and U.S. history ALS panelists were analyzed, along with the civics FT1 data for comparison. ACT was interested to learn how panelists reacted to various aspects of the field trials, and panelists' evaluations of the relative ease of rating items with the two methods was one important aspect to be determined.

Responses by panelists in civics FT1 using the ISSE method suggest that their understanding of the tasks, confidence in their ratings, and so forth were nearly as high as those reported by ALS panelists in geography and U.S. history after two rounds of ratings using the ME method in 1994. The responses to questions specifically about the rating methods were somewhat more positive for the ISSE panelists than had been the case for the panelists using the ME method when multiple-choice items were rated. The responses about using the ISSE method for rating constructed response items were somewhat less positive. The mean responses to those questions are reported in Table 7.

The writing FT1 data are reported in Table 8, and they are, of course, only for rating constructed response items. Those responses are mixed. Conceptual clarity was rated higher for panelists using the ME, but ease of application was rater higher for panelists using the ISSE method.

Table 7
Mean Response Score to Selected Questions about Methods in Round 2 Ratings
by Civics FT1 Panelists and ALS Panelists in Geography and U.S. History

	Civics FT1	Geography	U.S. History
The method for rating multiple-choice items was conceptually clear. (5 = Totally Agree; 1 = Totally Disagree)	4.63	4.32	4.43
The method for rating multiple-choice items was easy to apply. (5 = Totally Agree; 1 = Totally Disagree)	4.50	4.14	4.32
The method for rating constructed response items was conceptually clear. (5 = Totally Agree; 1 = Totally Disagree)	3.88	4.25	4.17
The method for rating constructed-response items was easy to apply. (5 = Totally Agree; 1 = Totally Disagree)	3.88	3.89	4.07

4 Responses to questionnaire items by panelists in each field trial are available upon request.

Table 8
Mean Response Score to Selected Questions about Methods in Round 2 Ratings
by Writing FT1 Panelists

	ISSE (FT1)	ME (FT1)
The method for rating prompts was conceptually clear. (5 = Totally Agree; 1 = Totally Disagree)	3.88	4.14
The method for rating prompts was easy to apply. (5 = Totally Agree; 1 = Totally Disagree)	3.75	3.00

Questionnaire data indicate that FT1 writing panelists using the ME method increased their confidence and understanding of the process and tasks more so than those using the ISSE method. Indeed, the conceptual clarity of the rating method (reported in Table 8 above for Round 2) increased from Round 1 to Round 2 for panelists using the ME method and the ME method became easier for them to apply. The responses of ISSE panelists indicated no improvement by Round 2 in the conceptual clarity of the rating method, and their evaluation of the ease of applying the method indicated that it was less easy to apply in Round 2 than in Round 1.

Consequences Data. Panelists were given consequences data based on their Round 2 ratings. An example of the format used for reporting consequences data is provided in Appendix 1. The percentages reported above in Table 3 for civics and Table 4 for writing are the consequences data shared with panelists in the field trials. Panelists were asked to evaluate the data and then they were asked to recommend new cutpoints if they felt the consequences data were not reasonable. The general changes recommended were reported in the **Process** section above.

The new cutpoints recommended by each panelist were averaged. For panelists who chose to recommend unchanged cutpoints, the grade level cutpoints from Round 2 were the values used to compute the new average. For civics, the averages are 150.78 for Basic, 170.24 for Proficient, and 186.48 for Advanced. Their recommendations were generally to lower the cutscores for each level. During the discussion they decided that those averages were to be their final recommendation. Only one person suggested a change from that average. Based on those new cutpoints, 64.3% of grade 8 students would score at or above Basic, 13.6% at or above Proficient, and 0.4% at or above Advanced.

For writing, there were many more changes recommended, as noted earlier. The same procedure was used for computing the new cutscores, based on the recommendations for change. The two groups were each engaged in a separate discussion of consequences data. Panelists using the ME method lowered the cutpoints for the Basic level somewhat, and they lowered the cutpoints considerably for both Proficient and Advanced. Their recommendations, based on the consequences data, resulted in 83.3% of the students scoring at or above the Basic level, 14.9% at or above the Proficient level, and .009% at or above the Advanced level. The Advanced cutscore was lowered from 213 to 206, but this was not nearly low enough to include even 1% of the grade 8 scores for students in the 1992 Writing NAEP.

For FT1 writing panelists using the ISSE method, only one panelist recommended changes in the Basic and Proficient cutpoints, and those changes had little impact on the final results. Five panelists recommended that the Advanced cutpoint be lowered, although two of those recommendations were for only very minor changes. The final Advanced cutpoint for the ISSE group was 217.3, and the percentage of students of student scores at or above this level was less than 0.00%.

Panelists were generally favorable to having consequences data and to having the opportunity to recommend changes. They felt, however, that recommending changes at the end of the rating process introduced arbitrariness to the process that caused them to feel less inclined to make changes and less positive about the opportunity. This reaction indicated that they would prefer having the consequences data earlier in the process, during the rounds of ratings.

Conclusions for Field Trial 1

Results from FT1 were to have determined which rating method to use in the ALS studies. ACT proposed that an item mapping method also be used in conjunction with whatever rating method was selected as a result of FT1. FT2 was then to be conducted to examine the interface of the two methods when used jointly, to evaluate how panelists perceived the two methods when used together and separately, and to determine the impact on cutscores using two rating methods in one process.

After carefully reviewing the data reported by ACT from the first field trial for each subject, TACSS was unable to recommend one of the two methods as the unambiguous choice to carry forward to the second field trial and remainder of the ALS process. Both NAGB and TACSS still wanted to have information from other methods in writing. TACSS recommended that ACT design the second field trial for civics with the two item-by-item rating methods used in field trial 1, and include the other research factors originally planned with one method. For field trial 2 in writing, they were eager to see results from several alternatives.

Detection of Bias in ISSE

Results of the field trials were viewed somewhat skeptically by TACSS. There was concern that the ISSE method would result in more extreme cutscores due to a bias. That is, some TACSS members felt that the method would necessarily result in lower Basic cutscores and higher Advanced cutscores, compared to the "true score" (Reckase, 1998; Bay, 1998). As a result of work by both Reckase and Forsyth, and based on the findings of the first field trials, TACSS recommended in May, 1998 that ACT discontinue further research using the ISSE method. They recommended that ACT explore alternatives already under consideration.

Field Trial 2: A Comparison of Methods and Timing of Consequences Feedback

ACT's plan had been to use the method selected on the basis of field trial 1 as the first of two methods used in setting achievement levels in field trial 2. The second method was to be an item mapping method that would allow panelists to make adjustments to their cutpoints directly. That is, rather than continuing with adjustments to item ratings after two or three rounds, panelists would switch to maps or charts with the items arrayed according to some statistical criteria, e.g., response probability=65%. The second field trial had been planned as an opportunity to study the interface of the two methods for setting achievement levels: an item-by-item rating method and an item mapping method.

Since no method had been selected as a result of the first field trials, and since the ISSE method had been eliminated from further consideration, the design of the second field trials had to be changed. The goal was still to test several alternatives for writing. Meanwhile, Reckase had proposed a new method, and TACSS agreed that it should be tested in the second field trials.

In addition to studying alternative standard setting methods—including interfacing methods, ACT wanted to collect data on the issue of providing consequences data. Specifically, ACT wanted data to help decide when in the process to provide the data and how often to provide it. When panelists are given consequences data at the end of the process, few recommend changes in response to the

data. NAGB needed to know whether this general trend would hold if panelists were given consequences data during the process when the cutscores could actually be altered. The data from FT1 would be supplemented through FT2 research.

ACT was to develop several design alternatives for consideration by TACSS. The designs from which TACSS selected are included in Appendix 2 along with the design adopted for writing FT2. Of those designs considered for civics FT2, Design 3 is the one adopted.

TACSS recommended that ACT test the Mean Estimation Method with Item Mapping and the Mark Reckase Method in the civics FT2. In order to include research on consequences data, four groups were needed for the study, and each group was to include 10 panelists each.

The decision on methods to test in the writing FT2 was more difficult. TACSS was somewhat divided with respect to the choice. In order to collect data on the impact of consequences data in writing FT2, no more than two different rating methods seemed feasible. Ultimately, TACSS decided that the Booklet Classification method and the Mark Reckase methods should be tried in this study. The Grid method was rejected because the computational procedures had not yet been determined and it was a totally new procedure (Bay and Loomis, 1998). The two methods selected included both an item-by-item and a holistic approach. Before describing the process, more information about the methods is needed.

Alternative Methods for FT2

Item Mapping Method

The Item Mapping (IM) method investigated in the civics field trial is very similar to the Bookmark method used by CTB-McGraw Hill (Lewis, Mitzel, and Green, 1996). ACT implemented IM procedures in the 1996 Science NAEP ALS process (ACT, 1997c), for research purposes and for grade 8 panelists when they were reconvened and given to change their ALS recommendations. The IM method uses a linear chart that indicates the approximate range of test scores earned on the NAEP. Each item was located or mapped at a point on the ACT NAEP-like scale where student performance reaches a 65% probability of correct response for the item. By studying the item map, one is able to determine which test items were responded to correctly 65% of the time by students who scored within a certain range (i.e., achievement level) on the score scale. Items were identified by a sequential number representing their rank with respect to difficulty. Abbreviated item descriptions, along with the score point at which each item "mapped" were included in the materials provided to panelists.

Dichotomous items are mapped directly. Polytomous items are dichotomized at each score level and then mapped to the scale in the same manner as dichotomous items. Thus, each polytomous item is mapped to the scale one time fewer than the number of score levels. FT2 used a similar mapping procedure. The exact mapping criteria were never fully "approved" by TACSS because they were never fully in agreement with the use of item maps. Ultimately, the decision was to use the criterion used most frequently in NAEP reporting, i.e., 65%.

The IM method was studied in conjunction with the ME method. Panelists in the ME group rated items on an item-by-item basis for two rounds, and then they were given item maps and lists with item information to identify the items. They examined the maps and their other feedback data to decide whether the group cutscore should be modified. They recorded the recommended cutpoint for each level on their item map. The recommended cutpoints were averaged for computing the final cutpoint for each rating group.

The Mark Reckase Method

ACT had experimented over the years with various methods of providing information about intrarater consistency. TACSS had generally judged the efforts as less than successful, and intrarater consistency feedback was dropped from the feedback provided to panelists in the 1996 Science ALS process. The Reckase method addressed the need to provide panelists with useful, easily understood information. The information in the Reckase Charts would help them evaluate the consistency of their ratings for items of different formats, different content dimensions, and other item characteristics.

Reckase proposed to have panelists use an item-by-item rating method to generate an initial set of ratings. The modified Angoff/ME method was suggested. Charts, now known as "Reckase Charts," were presented to panelists after the first round of item ratings. These Reckase Charts include expected student response scores for each item at each point on the score scale. A column on the chart contains expected score data for one item across all score points. A row contains expected score data for one point on the score scale across all items. The expected score data are generated by the IRT model. For polytomous items, the expected score data are reported as a mean score for each item. For dichotomous items, the expected score data are reported as the probability of correct response/percentage of students responding correctly.

For writing, all prompts in the rating pool could be printed on one large chart. For civics, each block required a separate page or chart. For FT2, separate color-coded charts were prepared for each of the three achievement levels: **B**asic ratings were marked on **b**lue charts, **P**roficient ratings on **p**ink, and **A**dvanced on **a**methyst.

The Reckase Method required that panelists transfer their item ratings from Round 1, for example, to the charts. By marking ratings on the charts, panelists would be able to visually inspect their ratings for each item with respect to their own individual cutscore, the grade level cutscore, item type (multiple choice and constructed response), and content/prompt type (persuasive, informative, and narrative, for example).

Panelists inspected the charts, along with other feedback data, and decided on ratings for each item in a second round of item-by-item ratings. The third round of ratings required panelists to select a row, i.e., a score, to represent their cutpoint for each achievement level.

The amount of information available to panelists through the use of Reckase Charts was great. There was some concern, however, that ratings would be "data driven," and that panelists would loose their focus on the achievement levels-descriptions (the standards) to be used in making their judgments. As a result of this concern, TACSS recommended that the charts presented after the first round of ratings exclude the ACT NAEP-Like scores. Each row on the charts was identified with alpha coding. Panelists would only have their item ratings on the charts to evaluate before round 2 ratings. They were instructed to examine items for which ratings appeared particularly high or low to determine whether any patterns emerged based on item type or content area, and to pay particular attention to ratings for which their confidence was especially high or low when rating the item. An example of a chart is included in Appendix 2.

After Round 2, new charts were again distributed to panelists, and they again transferred their ratings to the charts. This time, the ACT NAEP-Like score points were included on the charts. Panelists could evaluate their ratings relative to the grade level cutscores and relative to their own cutscores although they were not explicitly instructed to do so.

The Reckase Method was to be tested in the second field trial for each subject. Because the final round of ratings was not an item-by-item rating procedure, item maps would not be tested in conjunction with the Reckase Method. Instead, Round 3 ratings in the Reckase Method required panelists to draw a line to identify the cutscore for each level. A space on the rating form was provided to record the cutscore for each level.

Booklet Classification

ACT had used a booklet classification method in validation studies for geography, U.S. History, and science (ACT 1995; 1997d). Results of those studies suggested that the cutpoints resulting from a Booklet Classification (BC) method would be higher than with the mean estimation method. ACT (Hanson, Bay, and Loomis, 1998) conducted further research on this method. Evidence suggested consistently that the cutpoints would be set higher with a BC method than with the ME method. Since "reasonable" outcomes are a goal of the ALS process, the method seemed to hold little promise. In addition to the reasonableness of outcomes, however, was the issue of how to compute cutscores with a BC method in NAEP. Alternatives were presented to TACSS for their review (Hanson, 1998).

There were many important issues to be resolved regarding the BC method (Bay, 1998). The number of booklets to be classified by panelists, the number of categories for the classification, the distribution of scores for booklets selected for the study and the criteria for determining the "score" in the NAEP context of plausible values, the number of different booklet forms to use to represent the assessment pool and not overburden the panelists, and so forth. Bay designed the study (Bay, 1998) and a more detailed description of the design implemented in the writing FT2 study is included in Hanson and Bay (1999).

Each panelist classified 40 booklets. There were 20 booklets in each of two different forms. Each panelist had forms including at least one prompt for each type of writing. In order to provide panelists the opportunity to discuss booklet classifications after Round 1, the design provided 10 booklets in each form to be classified by two people who would be seated together. Thus, each panelist had 10 booklets of form A to discuss with panelist A (on the right) and 10 booklet of form B to discuss with panelist B (on the left).

TACSS recommended that booklets be ordered on performance from lowest to highest. Panelists were told that the ordering was to facilitate their task and that it represented only one of many such orderings that might be used. They were told that their classifications did not have to reflect the ordering because classifications were to be made on the basis of the achievement levels descriptions.

Mean Estimation

The mean estimation method is the method used by ACT for setting achievement levels for NAEP since 1994. The method uses a modified-Angoff rating judgment for dichotomous items and estimation of the mean score for polytomous items (ACT, 1997c). The method was used in FT1 for both civics and writing.

The Panels

The plan for recruiting panelists for field trial 2 was the same as that planned for FT1. Given the lack of success with recruiting panelists in FT1 for each subject, however, the plan clearly had to change. TACSS advised that it was imperative that FT2 in each subject include at least 10 panelists for each method/procedure tested. In order to meet the requirements for the number of panelists, ACT scheduled the second field trials in the summer, after the school term had ended. The meetings were also scheduled on weekdays because of suggestions from nominators and

potential panelists that this would increase participation. Further, at the recommendation of TACSS, NAGB authorized ACT to offer an honorarium of $300 for the two-day field trials. ACT had no problem with recruiting the required number of panelists when these changes were announced to nominators. We accepted panelists who volunteered; our *usual* selection process was not implemented. We did attempt to recruit panels representing educators and non-educators. People in specific positions nominated candidates to serve on the field trial panel, and candidates were screened to assure that they had content knowledge and familiarity with students at grade 8.

For FT2 in civics, there were 43 panelists: 32 teachers, 5 nonteacher educators, and 6 general public members of the panels. There were 27 men and 16 women in FT2 for civics. For writing, there were 40 panelists: 30 teachers, 3 nonteacher educators, and 7 general public panelists. Thirty-three panelists in FT2 for writing were women and 7 were men. Panelists were assigned to the rating groups so that each was as equivalent as possible.

One error occurred in the civics FT2 assignment of panelists to groups. The initial assignments were made so that each rating method group was as equivalent as possible and, within each method group, each consequences data treatment group was also as equivalent as possible. During training exercises, however, staff noted that one particular group was taking far longer to complete tasks than others. The decision was made to reassign some panelists from one table to another, i.e., one consequences data treatment group to another, within the ME rating group. The reassignments were made on the basis of the panelist identification number. This reassignment made it much easier to distribute materials, but it resulted in having only teachers at one table rather than a mix of panelist types.

The Process

The Design

FT2 in each subject included two methods. For civics FT2, the Mark Reckase method and the ME method with item mapping were implemented. At least 10 panelists were assigned to each group. These are the four groups.

Civics FT2

a. mean estimation with item maps and consequences data after each round
b. mean estimation with item maps and consequences data after round 3
c. Reckase method with consequences data after each round
d. Reckase method with consequences data after round 3

Writing FT2

a. booklet classification with consequences data after each round
b. booklet classification with consequences data after round 2
c. Reckase method with consequences data after each round
d. Reckase method with consequences data after round 3

Implementation of the Process

The planned field trial process was implemented in each subject with relatively few problems. The same orientation and training were provided for the FT2 panels as were described previously for the FT1 panel. These panelists also took a form of the NAEP, just as all NAEP ALS panelists do. Participants were divided into equivalent groups, as described for FT1, and they were also assigned to table groups to be as equivalent as possible. The civics groups rated the same set of items described for FT1, and both methods groups used the same rating method during the first round. Civics FT2 panelists were trained together through the first round of ratings.

Writing panelists using the Mark Reckase (MR) method rated the same items described for FT1. Panelists using the Booklet Classification (BC) method had fewer different prompts in their pool for classification, but they were from the 1992 Writing NAEP and one prompt of each type was included in the forms for classification.

Panelists were again engaged in training exercises to become familiar with the assessment and the achievement levels descriptions. The first round of ratings/classifications was scheduled at the end of the first day. The facilitator of each group provided training in the rating methods for FT2 writing.

Cutpoints and other feedback data were computed and produced to distribute to panelists at the start of the second day. The feedback described for FT1 was provided to all panelists in the second field trials. One exception was that BC panelists did not participate in the whole booklet exercise. In civics, all panelists were trained together in the feedback common to the two methods. Following the general session of training, panelists were provided information in each rating group regarding feedback specific to their method. Groups C and D were instructed in Reckase Charts. Four sets of data were prepared for distribution to FT2 panelists. Although all ratings in FT2 civics were based on the same method for Round 1, separate data reports were prepared. Panelists in each group would have separate reports in subsequent rounds, and it seemed a good idea to give them separate reports beginning with Round 1 results and feedback. Because the rating methods for the writing FT2 were so different from the start, feedback was computed for each group separately and training in the feedback was conducted separately for each method group.
(Please refer to the design of FT2 for writing in Appendix 2.

Panelists in groups A and C received consequences data after Round 1 and after each subsequent round. Panelists in these groups were trained in consequences data and provided the information. A form was distributed to each panelist in the two groups and they were asked to comment on the consequences data. For FT2 in civics, the facilitator forgot to distribute the questionnaire until after the panelists had merged back with their rating groups. The panelists were interrupted briefly and asked to complete the questionnaire. No major problem was apparent as a result of this error.

Results

General Overview

Panelists were generally receptive to each method. Panelists found the Reckase Charts very informative and interesting. Similarly, FT2 civics panelists were enthusiastic about the information about student performance information represented in the item maps. Each method, or combination of methods was of interest to ACT. This was our first experience with any of the three methods, as such, being tested in the field trials. Panelists seemed to have no problems with the item maps (civics only) and the Reckase Charts (both civics and writing). Having the booklets ordered, in writing FT2 seemed to sharply change the task from the validation studies using booklet classification implemented previously by ACT. Rather than placing booklets in categories of achievement, they simply wrote their classifications on the booklets and on their "rating" form. Panelists did not classify booklets according to the ordering. That is, they did classify some booklets with higher ranks at lower levels than others around the same rank, and they classified booklets from lower ranks at higher levels than others around the same rank. Forty booklets, ordered on performance, and involving only two forms, did not present a challenging task to the panelists. They really appreciated the opportunity of discussing booklet classifications, and panelists in other groups seemed to really enjoy the opportunity of discussing item maps and Reckase Charts. ACT was further convinced of the importance of providing time for panelists to discuss tasks among themselves.

Findings for the writing field trials require some caution with respect to comparisons of cutscores. As was true for FT1, writing FT2 panelists worked with the achievement levels descriptions developed for the 1998 ALS process. In general, the cutscores set by FT2 writing panelists in both methods groups were lower than those from FT1. The cutpoints set at the Proficient level, and the consequences data associated with those scores, were particularly uncommon relative to ALS results for other subjects. The percentage of students scoring at or above the Proficient level set by panelists in writing field trials were generally quite low. As discussed below, panelists in the MR group in writing FT2 lowered the Proficient cutscore after seeing consequences data and reversed this general finding.

Findings

Round 1 cutscores and feedback ratings for panelists in Group A of FT2 civics showed that several raters gave much lower borderline Basic ratings for items than others in the group. The facilitator discussed their ratings with these panelists and they indicated that they had "gotten off track" in their ratings. The group as a whole was advised on how to interpret the feedback data, in light of these errors. The decision was made to use the group mean to replace the cutscores of these panelists in analyses of results. Data reported in Appendix 2 show results both with and without outliers.

Results from FT2 for the two subjects were inconclusive regarding the effects of consequences data on the cutpoints. In civics FT2, the group A panelists using the ME method and receiving consequences data throughout the process set cutscores lower across all rounds, taken together, than those in group B using the ME method and receiving consequences data later in the process. Results for the group using the MR method were just the opposite. That is, the group C panelists who received consequences data first after Round 1 set their cutscores higher across the rounds, taken together, than the group D panelists who received consequences data later in the process. (Please tables and charts in Appendix 2.) Overall differences between cutpoints for groups A and B using the ME method in civics FT2 were significant. Overall differences in cutpoints for groups C and D using the MR method were not significantly different. The timing of consequences data did not appear to have an effect on the cutpoints for the MR method. Timing of consequences data did appear to have an effect on the cutpoints for the ME method. Panelists who received consequences data earlier in the process set lower cutscores.

The data in Table 9 (below) report results for civics FT2 by rounds of rating and method/consequences groups. Recall that all panelists used exactly the same rating method for Round 1 ratings. Panelists in group A recommended no changes in their cutscores after Round 3, thus the final cutpoint for each panelist were all the same.

For writing FT2, the two methods implemented were quite different. The BC method groups classified booklets two times and then discussed consequences data, whereas the MR method groups had two rounds of item-by-item ratings before deciding their cutpoint for each level on the Reckase Charts.

In general, panelists in the BC method group set cutpoints across the three rounds that did not differ significantly by the timing of consequences feedback data. Data reported in Table 10 below show that BC panelists who received consequences data set slightly higher cutscores than those who did not. On the other hand, panelists in writing FT2 using the MR method and receiving consequences feedback data throughout the process generally set higher cutscores. At the Advanced level, cutscores set by MR panelists were higher than those set by BC panelists, no matter when consequences data were introduced.

Table 9
Cutpoints, Standard Deviations and Percentages of Students Scoring
At or Above Each Achievement Level for FT2 Civics: By Group and Round of Rating

	A (n=10) ME/IM		B (n=11) ME/IM		C (n=11) ME/MR		D (n=11) ME/MR	
Level	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>
Round 1								
Basic	136.4 (25.5)	89.8%	146.5 (12.2)	***74.3%***	145.9 (9.3)	75.1%	138.6 (11.3)	***87.4%***
Proficient	157.6 (5.1)	46.3	164.5 (5.1)	***26.6***	163.2 (4.4)	29.7	162.5 (4.5)	***32.0***
Advanced	168.9 (5.8)	16.0	175.3 (4.1)	***6.0***	173.9 (3.6)	7.9	174.1 (5.0)	***7.4***
Round 2								
Basic	146.7 (7.8)	73.5	151.3 (5.1)	***63.5***	148.4 (7.4)	70.0	143.5 (8.6)	***79.5***
Proficient	161.0 (4.8)	35.8	166.6 (3.1)	***21.6***	165.6 (2.7)	23.5	162.5 (4.7)	***32.0***
Advanced	171.8 (5.9)	10.9	177.3 (2.7)	***4.0***	177.0 (3.1)	4.4	175.5 (4.2)	***6.0***
Round 3								
Basic	149.7 (4.1)	67.1	153.2 (3.4)	58.6	149.2 (6.0)	68.2	146.5 (8.6)	73.5
Proficient	163.2 (2.9)	29.2	167.1 (2.8)	19.6	164.3 (3.3)	26.6	163.2 (4.1)	29.7
Advanced	174.5 (4.3)	7.0	178.0 (2.3)	3.4	176.7 (4.8)	4.4	177.4 (5.1)	4.0
Final								
Basic	149.7 (0.0)	67.1	153.1 (2.0)	58.6	145.9 (3.8)	75.1	147.6 (4.0)	71.8
Proficient	163.2 (0.0)	29.2	167.2 (1.1)	19.6	163.7 (0.7)	28.8	162.6 (1.3)	32.0
Advanced	174.5 (0.0)	7.0	177.9 (1.1)	3.7	175.0 (1.9)	6.5	177.2 (0.7)	4.4

Note: Data printed in ***bold italics were not presented to panelists in the process.***

Table 10
Cutpoints, Standard Deviations and Percentages of Students Scoring
At or Above Each Achievement Level for FT2 Writing:
By Group (n=10 each) and Round of Rating

	A Booklet Classification Consequences all Rounds		B Booklet Classification Consequences after Round 2		C Reckase Method Consequences all Rounds		D Reckase Method Consequences after Round 3	
Level	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>
Round 1								
Basic	128.1 (7.6)	96.9%	129.7 (11.1)	***96.1%***	138.5 (14.6)	97.8%	123.1 (14.0)	***98.7%***
Proficient	157.1 (8.4)	44.8	160.6 (6.6)	***34.9***	179.4 (9.8)	3.6	164.6 (13.2)	***25.5***
Advanced	194.5 (11.8)	0.2	187.1 (14.0)	***0.8***	217.5 (13.1)	0.0	211.3 (13.7)	***0.0***
Round 2								
Basic	131.6 (8.9)	94.9%	131.5 (7.5)	94.9%	134.1 (20.1)	92.9%	124.8 (11.5)	***98.4%***
Proficient	156.8 (11.1)	44.8	154.3 (8.6)	52.3	167.5 (16.9)	19.0	166.3 (15.3)	***21.4***
Advanced	196.5 (9.7)	0.1	190.0 (8.9)	0.4	202.9 (16.1)	0.0	213.5 (11.3)	***0.0***
Round 3								
Basic	Not		Not		136.2 (11.4)	90.7%	122.6 (11.1)	98.8%
Proficient	Applicable		Applicable		171.5 (11.6)	11.9	165.7 (11.7)	22.3
Advanced					208.6 (18.0)	0.0	215.7 (14.9)	0.0
Final								
Basic	133.8 (4.2)	93.2%	131.8 (3.1)	94.7%	137.3 (2.8)	89.9%	125.4 (2.4)	98.1%
Proficient	157.6 (2.6)	42.8	156.3 (3.3)	47.0	164.9 (7.2)	24.8	154.4 (4.2)	53.4
Advanced	191.8 (3.9)	0.3	187.0 (3.8)	0.9	198.6 (9.3)	0.0	201.3 (12.1)	0.0

Note: Data printed in ***bold italics were not presented to panelists in the process.***

Panelists in the Booklet Classification method seemed to understand the consequences data less well than other panelists. They were confused by the results that reported essentially no students at or above their cutpoints at the Advanced level. This confusion resulted from the fact that they had classified booklets at the Advanced level. They found it hard to understand why the proportion of the booklets they classified as borderline Advanced and Advanced were not more nearly reflected by the consequences data.

The general pattern of change in cutpoints for the final round was to raise the Basic cutpoint slightly and lower the Advanced cutpoint considerably. Both groups of BC panelists raised the Proficient cutpoint, and both groups of MR panelists lowered the Proficient cutpoint. The MR group receiving consequences data for the first time after Round 3 lowered the Proficient cutpoint by 11 points and increased the percentage of students scoring at or above the cutpoint from 22% to 53%.

While all groups lowered the cutpoint for the final Advanced cutpoint, the MR group that first received consequences data after Round 3 lowered their cutpoint most. They lowered their cutpoint by 14.4 points. Even so, that score point was still generally well above the range of student performance on the 1992 NAEP.

Data in Table 10 above (and in figures in Appendix 2) show the standard deviations for ratings across the rounds by method and consequences feedback groups. The standard deviations were considerably lower for the BC groups than for the MR groups. Standard deviations for Round 2 were generally higher for groups receiving consequences feedback data prior to that round than for groups not receiving the data. The Reckase Charts in the MR method can reveal extensive information to panelists. Further, the MR method requires panelists to shift from an item-by-item rating method to a more holistic method of identifying a score point on the Reckase Chart. Perhaps that accounts for the relatively higher variability among raters. The Advanced cutpoints were sharply higher than for the other two achievement levels for group C panelists for Round 3 and the final cutpoints.

Evaluations by Panelists. Vast amounts of evaluation data were collected from the four groups of panelists in FT2 for the two subjects.[5] Data tables have not yet been prepared for the writing FT2 evaluation, however, and only data from the civics FT2 can be reported. (Please refer to tables in Appendix 2.) Further, only a few examples of results are presented here, along with a brief summary of the general findings. Data such as those reported below in Table 11 were analyzed across rounds between consequences groups within rating method group, and between rating method groups. We generally expect to find that methods become both clearer and easier to apply with each successive round of application, for example. Here, we see that panelists were somewhat less clear about the rating methods for multiple choice items in Round 3 than they had been in Round 2. The decline was not consistent across all groups for constructed response items. The same pattern is observed for the ease of applying the method for multiple choice items for panelists using the item maps in group A and for panelists marking their cutpoints on the Reckase Charts in group D.

Evaluations by panelists revealed somewhat less confidence in ratings and less understanding of the methods after three rounds among civics FT2 panelists who received consequences data than those who did not. Panelists in the MR method group who received consequences data throughout the process were less confident in their selection of a cutpoint for each achievement level than panelists who did not get the data until later, for example. The panelists receiving consequences data

5 Complete data for both subjects are/will be available upon request. The data are too extensive to reproduce in this report, but some tables are presented in Appendix 2 for civics FT2.

Table 11
Mean Response Score on 5-Point Likert-Type Scale to Selected Questions About Rating Methods for Civics FT2

	Group	Round 1	Round 2	Round 3
The method for rating multiple-choice items was conceptually clear. (5 = Totally Agree; 1 = Totally Disagree)	A	4.1	4.6	3.5
	B	4.2	4.6	4.4
	C	4.2	4.0	4.2
	D	4.6	4.4	4.0
The method for rating multiple-choice items was easy to apply. (5 = Totally Agree; 1 = Totally Disagree)	A	4.0	4.5	3.5
	B	3.7	4.4	4.3
	C	4.0	3.9	4.0
	D	3.9	4.3	4.0
The method for rating constructed response items was conceptually clear. (5 = Totally Agree; 1 = Totally Disagree)	A	3.2	4.4	3.5
	B	3.9	4.1	4.2
	C	3.6	3.8	3.9
	D	4.2	4.0	3.7
The method for rating constructed-response items was easy to apply. (5 = Totally Agree; 1 = Totally Disagree)	A	2.9	4.1	3.6
	B	3.6	3.7	4.2
	C	3.4	3.8	3.9
	D	3.6	3.8	3.7

throughout the process less frequently responded that the Reckase Charts were informative and revealing with respect to the consistency of their ratings on various dimensions related to item types. Relative to the other panelists in the MR group, panelists in group A found the Reckase Charts less helpful and less likely to bring their cutpoints closer to their concept of borderline performance for each level than their ratings had been without the data in the Reckase Charts. The differences in mean responses were not great, but this pattern was generally observed.

Similarly, the ME group receiving consequences data throughout the process was somewhat less positive in their responses about the process and their ratings than panelists receiving consequences data later. Again, the differences in the mean responses for the two groups were not great, but this pattern generally held. There was no obvious explanation for how consequences data, per se, could impact panelists' understanding of/ability to use/confidence in using item maps, for example. These differences seemed more attributable to the general "personalities" of the panelists than to the effects of consequences data, however.

These patterns could have been a result of information "overload." There was a general concern that panelists were given more information than they could absorb in such a short amount of time. The field trials lasted less than half the time devoted to ALS meetings.

Reactions to Consequences Data. Recall that statistically, the differences between civics FT2 cutpoints for the ME groups were not significant and they were for the MR groups. In general, the consequences data appeared to have little effect on the panelists in the ME group and to have a greater effect on the panelists in the MR group. This observation is based on the number of changes recommended by panelists in the two groups in response to the consequences data provided.

Panelists using the Reckase method tended to recommend more changes. The MR group receiving consequences data throughout the process recommended more changes after Round 2 than after Round 1. All, or almost all, panelists at one table in MR group A recommended changes to all three levels following their third review of consequences data. Seven panelists in the MR group receiving consequences data only after Round 3 recommended 11 changes in cutpoints, but only 7 changes were recommended by 3 panelists in the similar group using the ME method. Panelists in the ME group receiving consequences throughout the process made no changes after the second round of consequences data.

The results for writing FT2 were less clear. Panelists in the BC method were generally more confused by the consequences data than panelists in other groups had been. Perhaps the confusion resulted from the fact that they were classifying booklets into categories. They reasoned that "a real student wrote each booklet," and they expected some students in the Advanced level. They were reminded several times that their 40 booklets did not reflect the national distribution of student performance, and their comments suggested that they tried to keep this in mind. Still, they had difficulty reconciling the consequences data with their classifications. When asked to recommend final cutpoints, the panelists tended to recommend percentages within levels rather than percentages at or above levels or actual cutpoints, as requested on the consequences questionnaires.

No data tables for writing FT2 have been produced yet to show cutpoint changes for each panelist at each round by consequences treatment group. Only the overall analyses of cutpoint differences were conducted in the limited time between field trials and pilot studies. Those results seemed sufficient to show that the effect of timing and frequency of consequences data was not consistent across the methods.

Recommendations for Methods and Procedures to Implement in Pilot Studies

Materials from the field trials were presented to TACSS during four different two-day meetings that were held prior to the pilot studies. During their July 1998 meeting, TACSS recommended the methods and procedures to be used in the pilot studies. These methods were, of course, also to be implemented in the ALS meetings unless some evidence was revealed in the pilot studies to cause modifications.

One concern was that panelists in the field trials had been given too much feedback. TACSS urged ACT to plan carefully the feedback to be presented to panelists, the sequencing of the feedback, and the instructions. Panelists need time to think about the feedback before applying it.

Panelists in the field trials were not as carefully screened, as they will be for the pilots and ALS panels. And, there was far less time in the 2-day field trials than in the 5-day ALS meetings for panelists to absorb the information. Still, there was concern.

Consequences Data

The findings regarding the timing of consequences data were neither conclusive nor unexpected. TACSS has consistently recommended that panelists be informed about the consequences of their judgments. TACSS was not, however, convinced that the information would have a great impact on subsequent judgments of panelists. TACSS simply believed that panelists should have the data. Recognizing that NAGB has never approved the use of consequences data within the ALS process, however, TACSS recommended that the consequences data be provided for the first time after Round 3. That gives panelists the opportunity of recommending modifications to the cutpoints after seeing the consequences data and the recommendations of panelists will be used to compute the final cutpoints. The final cutpoints will be recommended to NAGB, unless there was a reason not to

do so. The final cutpoints will also be used in the selection of exemplar items and performances to be used in reporting student performance relative to the achievement levels.

Following this recommendation, cutpoints were to be produced from the third round of item ratings without consequences data and from cutpoints based on modifications to those Round 3 cutpoints made in response to consequences data. TACSS recommended that NAGB be provided with both Round 3 and Final cutpoints.

TACSS also recommended ACT provide individual level consequences data to panelists after Round 3. They reasoned that panelists should have the opportunity of adjusting their own cutpoints, based on data about the consequences of those cutpoints. They recommended that rater location charts be modified for Round 3 to include data about the proportional distribution of student scores. These charts would provide a visual representation of their own cutpoints and help panelists determine whether, in what direction, and by how much to modify their cutpoints. TACSS also recommended that panelists be provided with data reporting the cutpoints and consequences data for each panelist (using codes for identification) in their grade group. These modifications would provide panelists with more information to use in deciding on their final cutpoints.

Rating Method(s)

TACSS generally found no compelling reason to choose one method over another, based on field trial data alone. They were interested in how well panelists seemed to understand the process, and they looked for any indications that one method would produce more reasonable, consistent, reliable results. They placed a high value selecting a method for which considerable research had been conducted and for which ACT had relatively more experience. This pointed to the choice of the ME method, i.e., modified Angoff for dichotomous item ratings and estimation of the average score for polytomous items.

TACSS did not, however, recommend the use of the ME method with item maps. Although ACT has conducted several different research studies with different mapping criteria, the choice of a response probability for mapping items remains an unresolved issue. The response probability (RP) used for mapping items determines the actual cutpoint, and the choice is clearly significant. In the absence of a policy regarding the RP value to use, TACSS recommended against the use of an item mapping procedure.

Both ACT Project Staff and TACSS were impressed with the apparent ease with which panelists used the Reckase Charts, and they believed that the information available to panelists through the Reckase Charts should be incorporated into the process. Yet, there was concern that the MR method held the potential for being too "data driven" and that the final cutpoints would be based on chart data rather than the standards.

TACSS reviewed the results of the Booklet Classification method and additional analyses by ACT of cutpoints based on borderline booklets versus cutpoints based on all booklets classified at the borderline and within the levels. The differences were disturbing. TACSS discussed alternative computational methods, but they decided to recommend against the use of the BC method for writing. This decision was based on their concerns about the computational procedures for the BC method in the NAEP context. It was also based on their concerns regarding the extensive production and logistics requirements associated with the BC method.

The recommendation was to use the Mark Reckase method, but NOT have panelists identify cutpoints on the charts for the final round. That is, have panelists use the ME method for rating

items through three rounds. The Reckase Charts will be provided to panelists prior to Round 2 and Round 3 ratings to inform them and to help them decide whether and how to modify ratings.

They also recommended that the ACT NAEP-Like scale scores be printed on the Reckase Charts for each round. There was discussion regarding the possibility of having ratings marked electronically on the charts, but ACT Project Staff voiced doubt that there would be enough time to perform this task between Rounds 2 and 3 which occur within a few hours of each other. Further, some still believed that the panelists would gain a fuller understanding of their ratings if they marked them on the charts.

TACSS also recommended that panelists be instructed to draw lines on each Reckase Chart to represent both their own cutpoint and the grade cutpoint for each achievement level. This would allow panelists to examine their ratings with respect to these cutpoints. TACSS suggested explicit instructions for panelists regarding the interpretation of the data on the charts, relative to cutpoint data.

Summary

The field trials were conducted to test rating methods and the impact of consequences feedback. The field trials provided the opportunity to try out different methods similar to those used successfully by others, as well as to try out some new methods. These field trials contribute significantly to the advancement of research information regarding some alternative standard setting methods.

ACT had proposed a new method to be tested in the field trials, once it "passed the test" in simulation studies. Although successful implementation of the method (or a very similar version) had been reported, that method was found to be biased, and ACT stopped tests with the method after the first field trials.

Reservations about the use of item maps were not overcome in the field trial process, and item maps were eliminated as a choice. Concerns about computational procedures and about the logistic demands of the Booklet Classification method eliminated this method

TACSS strongly recommended that the method used for the 1998 ALS process have a solid research base. TACSS had found no real reason to change methods and would not recommend doing so unless the alternative offered significant potential improvements in the process.

TACSS recommended a new "combination" method combining the greatest benefits of the new Mark Reckase method with the strong research base and extensive experience by ACT associated with the Mean Estimation method. The recommendation proved to be a good one, and the procedures were implemented successfully to set achievement levels for the 1998 NAEP in Civics and in Writing.

References

ACT (1995). *Preliminary report on the 1994 National Assessment of Educational Progress achievement levels-setting process for U.S. history, and geography.* Iowa City, IA: Author.

ACT (1997a). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume 1. Pilot study 1.* Iowa City, IA: Author.

ACT (1997b). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume II. Pilot study 2.* Iowa City, IA: Author.

ACT (1997c). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume III. Achievement levels-setting study.* Iowa City, IA: Author.

ACT (1997d). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science: Final report, Volume IV.* Validity evidence and special studies. Iowa City, IA: Author.

ACT (1997e). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science. Final report, Volume V. Technical decisions and NAGB actions.* Iowa City, IA: Author.

ACT (1997f). *Developing achievement levels on the 1998 NAEP in civics and writing: Technical proposal.* Iowa City, IA: Author.

ACT (1997g). *Developing achievement levels on the 1998 NAEP in civics and writing: Design document.* Iowa City, IA: Author.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement (2nd Ed.).* Washington, DC: American Council on Education.

Bay, L. & Loomis, S.C. (1998). *Setting achievement levels cutpoints using the grid method.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

Bay, L. (1998). *Booklet classification method: The issue of booklets to be classified.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

Bay, L. & Hanson, B. (1997). *Computing achievement levels cutpoints from NAEP BCS: A Secondary Analysis.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

Carlson, J. (1995). *Estimation of Reliability of NAEP IRT Proficiency Score Estimates.* Technical Memorandum, ETS.

Chen, L. (1998). *Setting achievement levels standards using item score judgment: simulation study.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

Hambleton, R. K. & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-56.

Hanson, B.A. & Bay, L. (1999). *Classifying student performance as a method for setting achievement levels for NAEP writing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Hanson, B.A. (1998). *Application of cubic regression method using booklet classification data.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

Hanson, B.A., Bay, L. & Loomis, S.C. (1998) *Booklet classification method.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

Impara, J. C., & Plake, B. S. (1996). *Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method.* Paper presented at the annual meeting of the National Council of Measurement in Education, New York.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring.* Symposium presented at the 1996 Council of Chief State School Offices 1996 National Conference on Large-Scale Assessment, Phoenix, AZ.

Loomis, S.C. (1998). *Summary of civics field trial #2.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

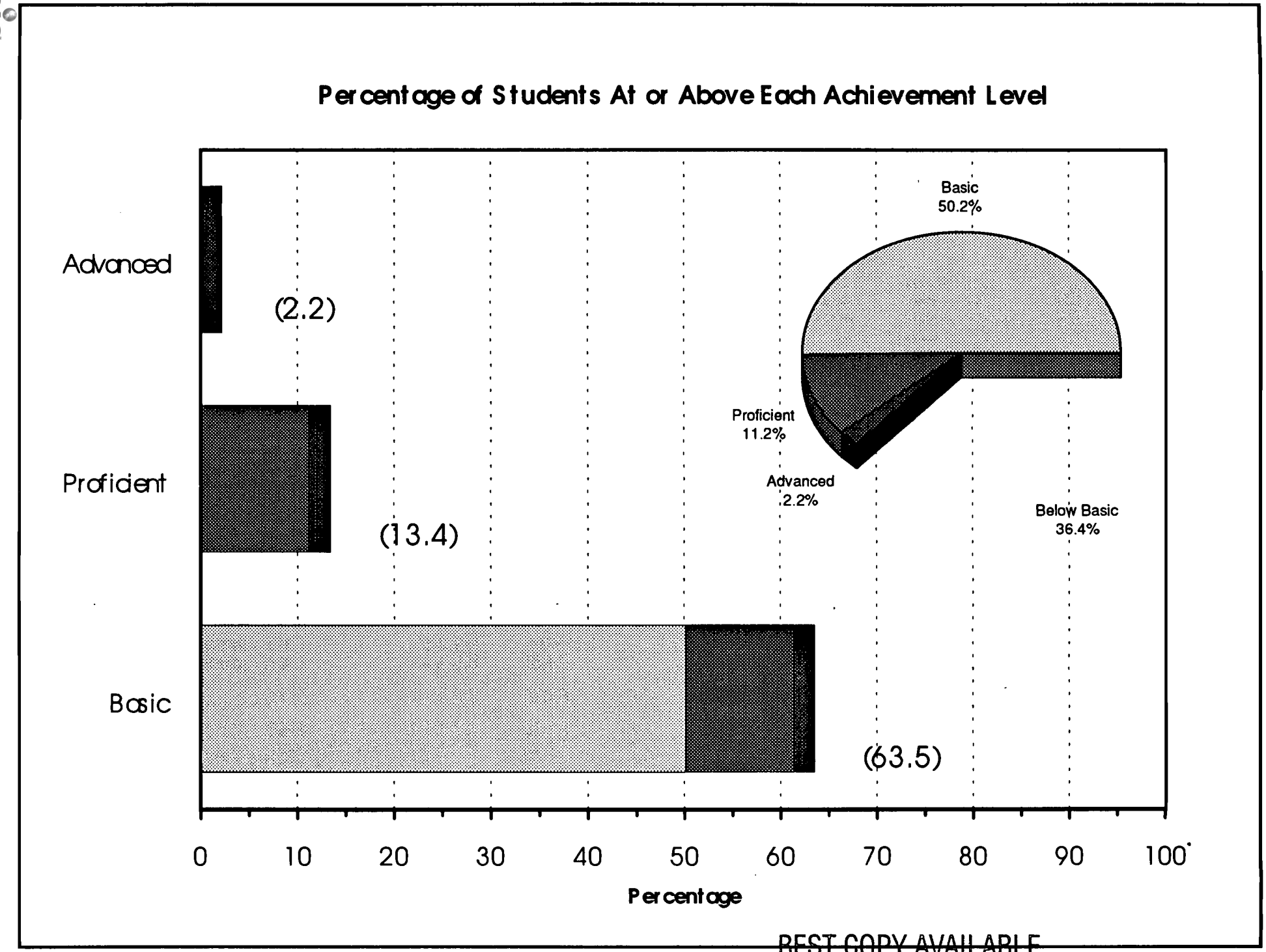
National Academy of Education (1993). *Setting Performance Standards for Student Achievement*, Robert Glaser, Robert Linn, and George Bohrnstedt, eds. Panel on the Evaluation of the NAEP Trial State Assessment. Stanford, CA: Author.

National Research Council (1999). *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress*, James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell, eds. Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment. Washington, DC: National Academy Press.

Reckase, M.D. (1998). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale: The 1996 science NAEP process. *Applied Measurement in Education , 11,* (1): 9-21.

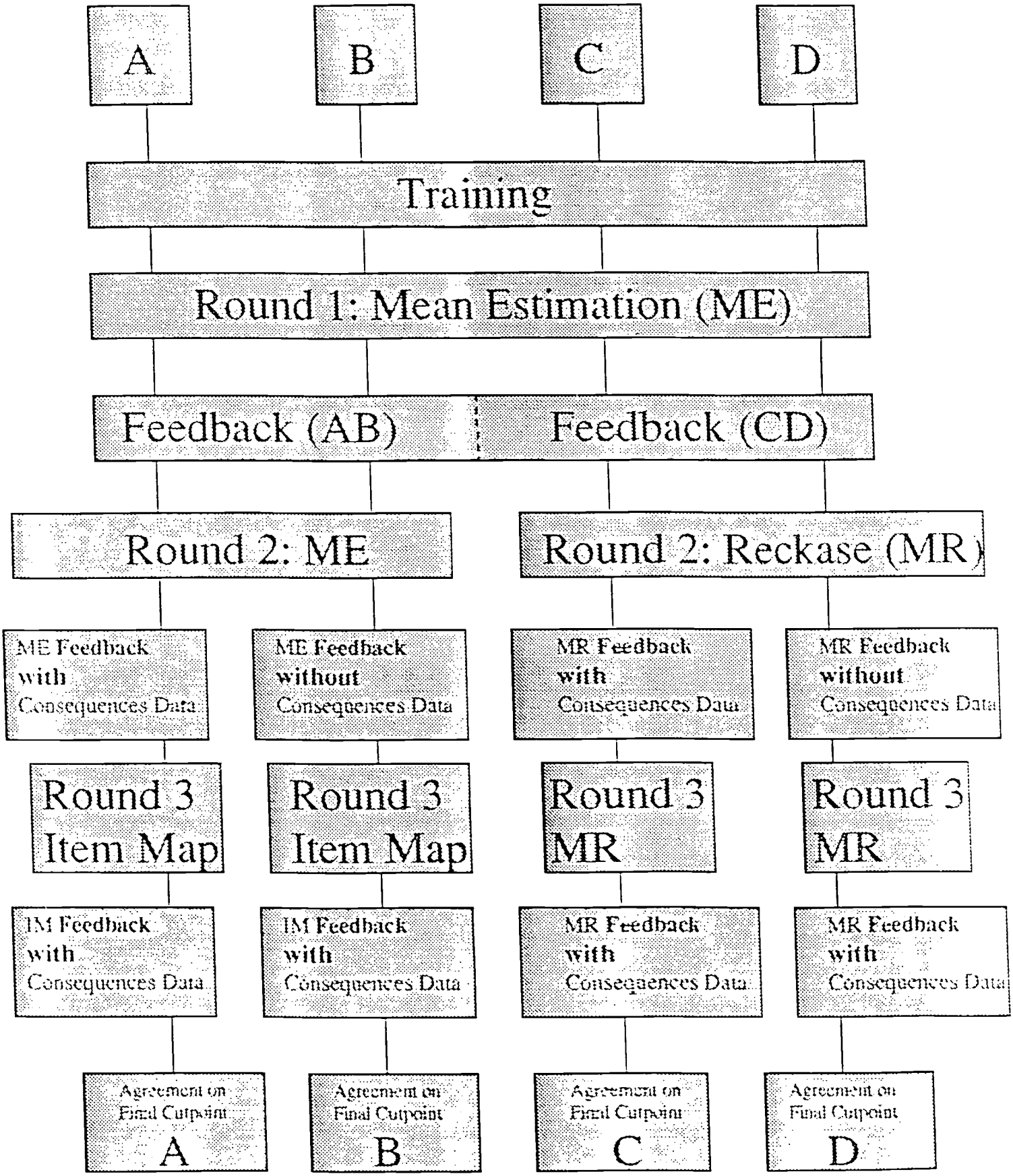
Shepard, L.A. (1995). *Implications for Standard Setting of the NAE Evaluation of NAEP Achievement Levels.* Proceeding of the Joint Conference on Standard Setting for Large Scale Assessments. Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Appendix 1

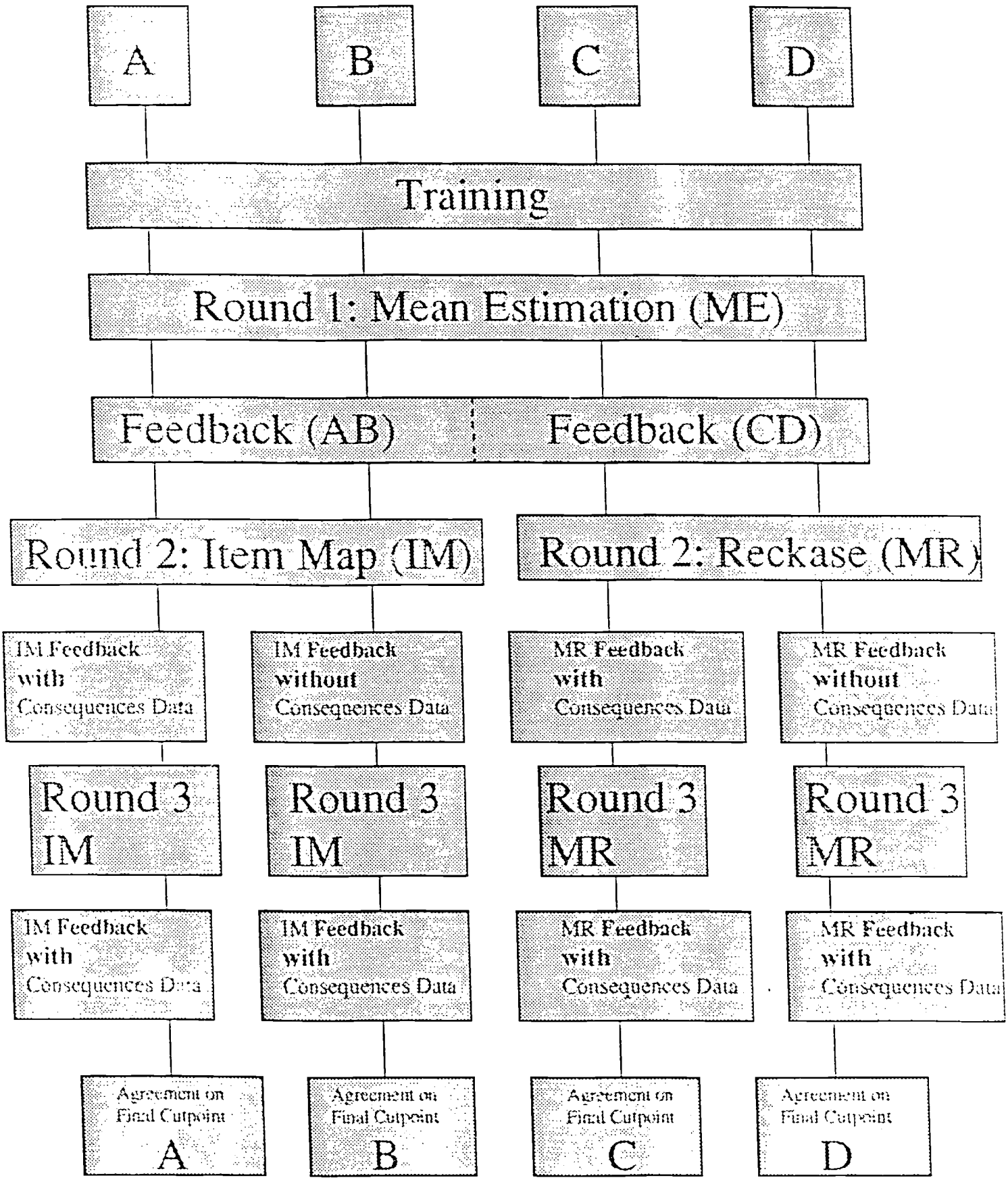


Appendix 2

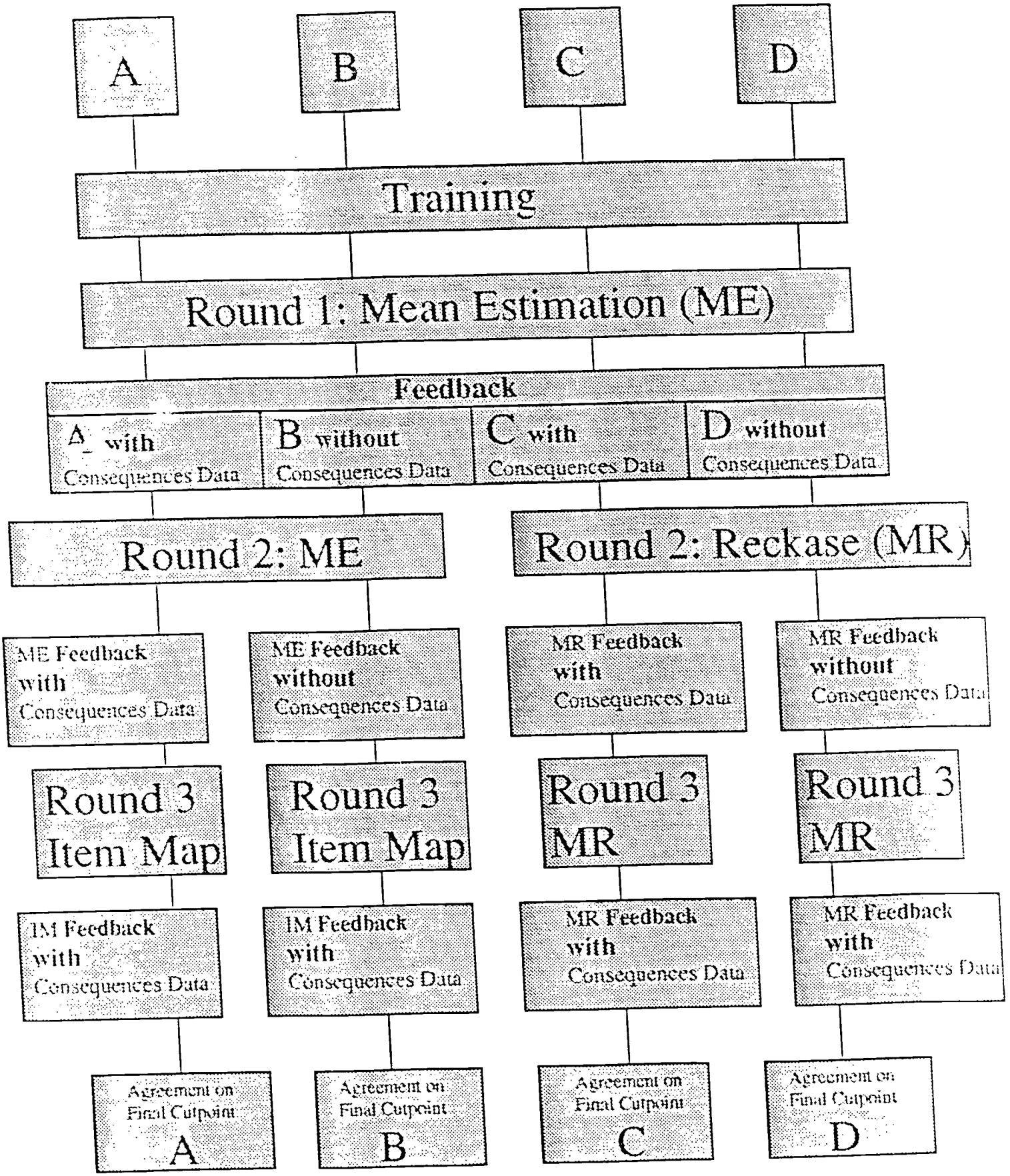
FT2 Design 1



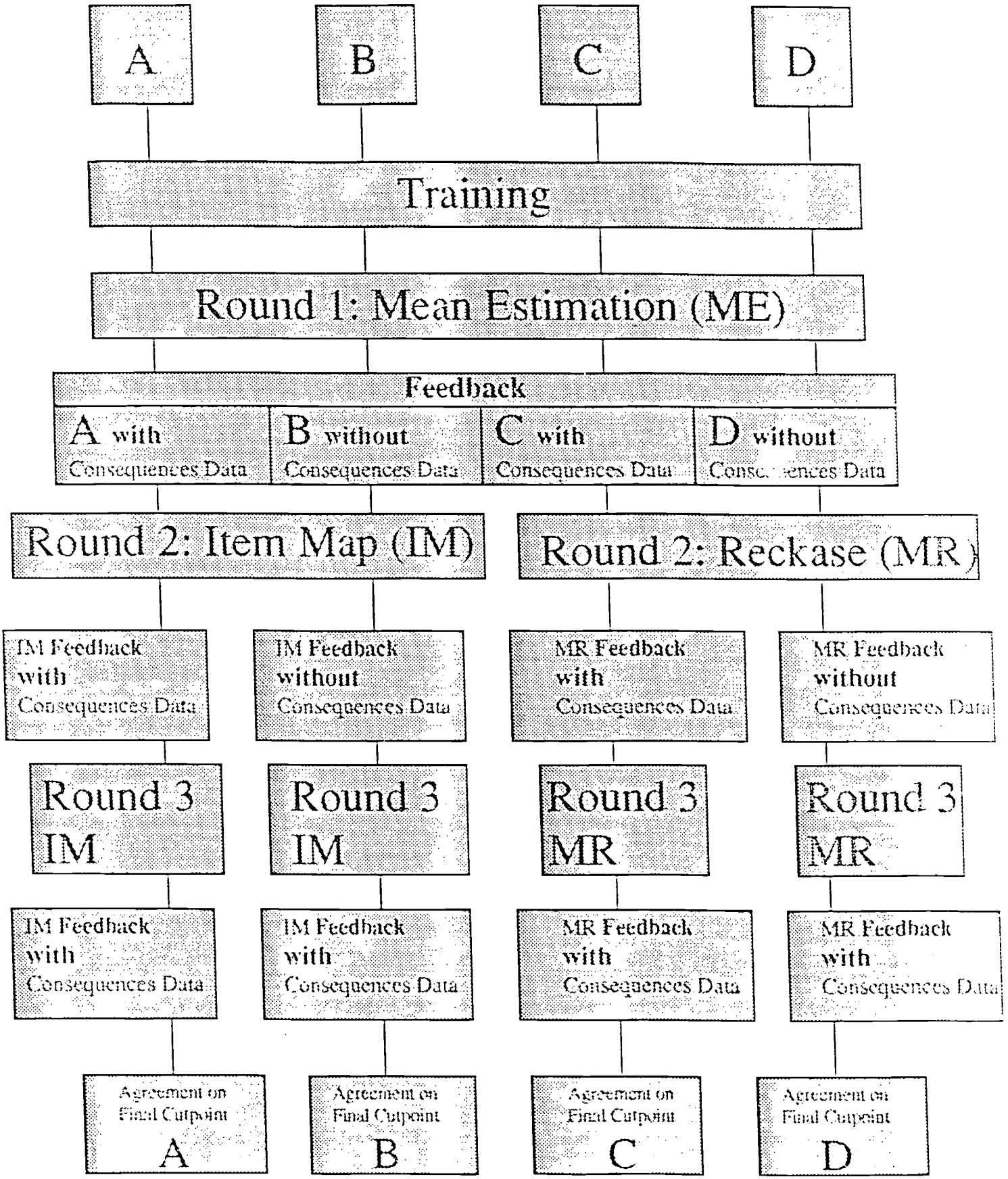
FT2 Design 2



FT2 Design 3



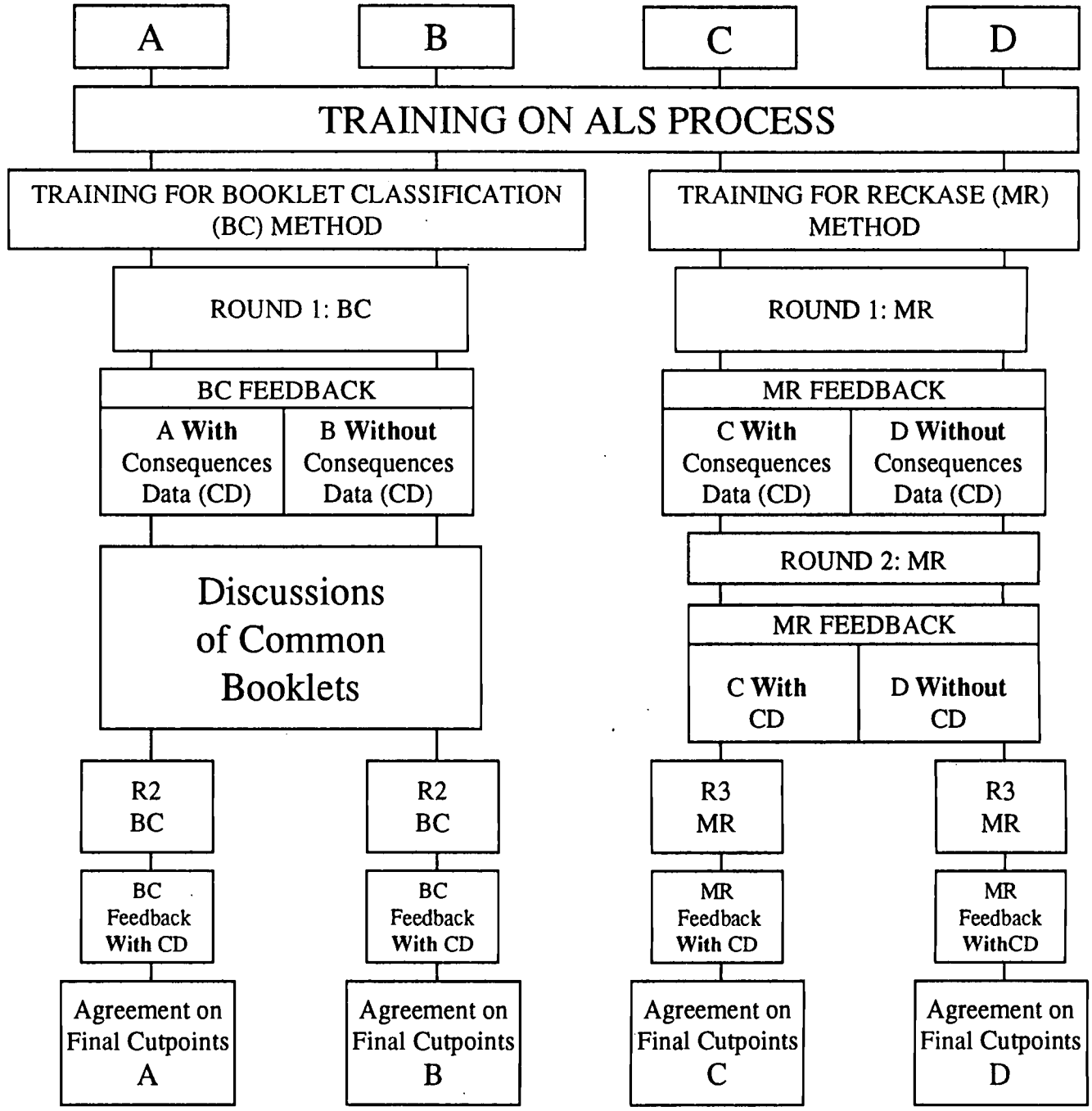
FT2 Design 4



1998 NAEP

Achievement Levels-Setting Process

Field Trial 2 for Writing



Civics Field Trial 2
Descriptive Statistics—Outliers Replaced with Means

Means and Standard Deviations Across All Groups and Rounds

Variable	N	Mean	SD
Basic	129	148.5	7.6
Proficient	129	164.2	4.5
Advanced	129	175.8	4.6

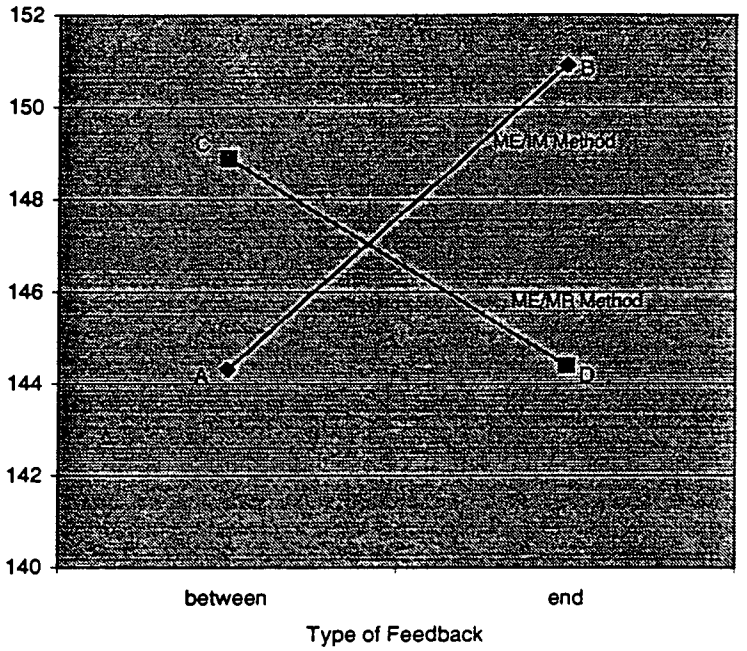
Means and Standard Deviations by Rounds Across All Groups

Round	Variable	N	Mean	SD
1	Basic	43	147.7	8.7
	Proficient	43	163.8	5.4
	Advanced	43	175.0	5.1
2	Basic	43	148.0	7.6
	Proficient	43	164.2	4.4
	Advanced	43	175.8	4.4
3	Basic	43	149.7	6.2
	Proficient	43	164.5	3.6
	Advanced	43	176.7	4.3

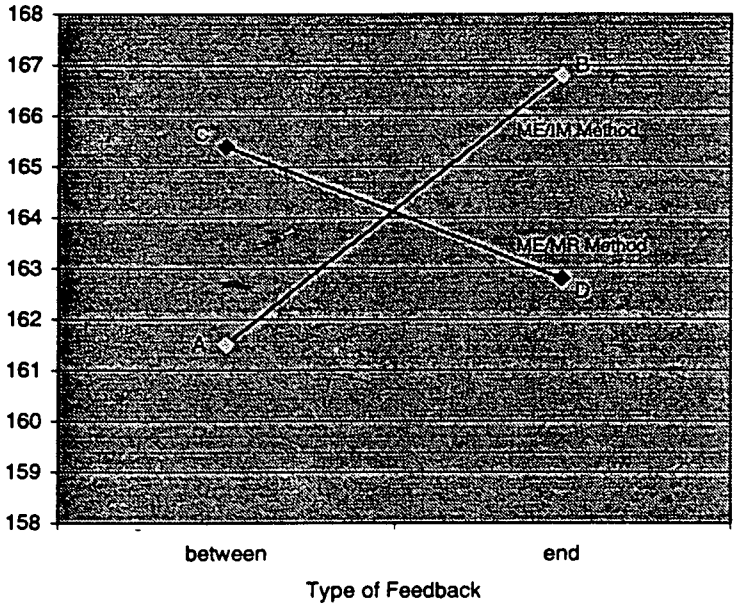
Means and Standard Deviations by Groups Across All Rounds

Group	Variable	N	Mean	SD
A	Basic	30	148.6	5.6
	Proficient	30	161.5	4.4
	Advanced	30	172.6	5.4
B	Basic	33	151.9	5.0
	Proficient	33	166.8	3.7
	Advanced	33	177.5	3.0
C	Basic	33	148.9	7.5
	Proficient	33	165.4	3.5
	Advanced	33	176.8	3.8
D	Basic	33	144.4	9.4
	Proficient	33	162.8	4.3
	Advanced	33	176.0	4.8

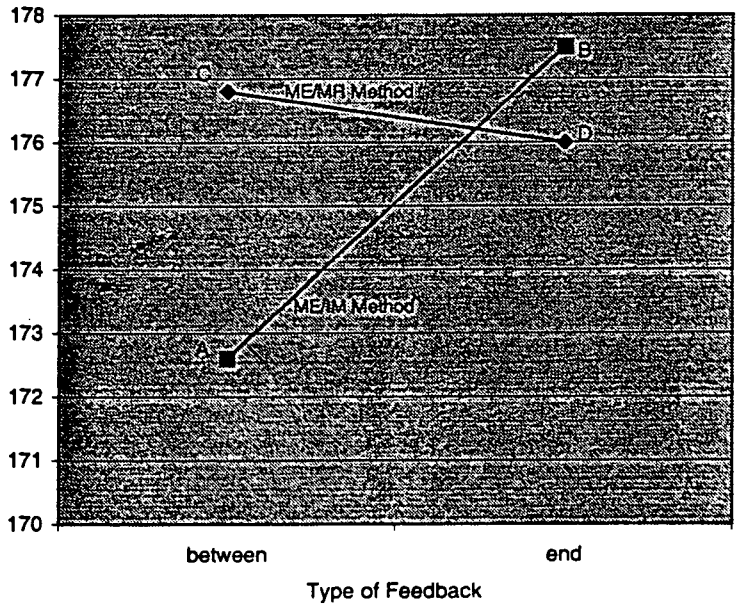
Mean Basic Cutpoints for Rating Method by Type of Feedback



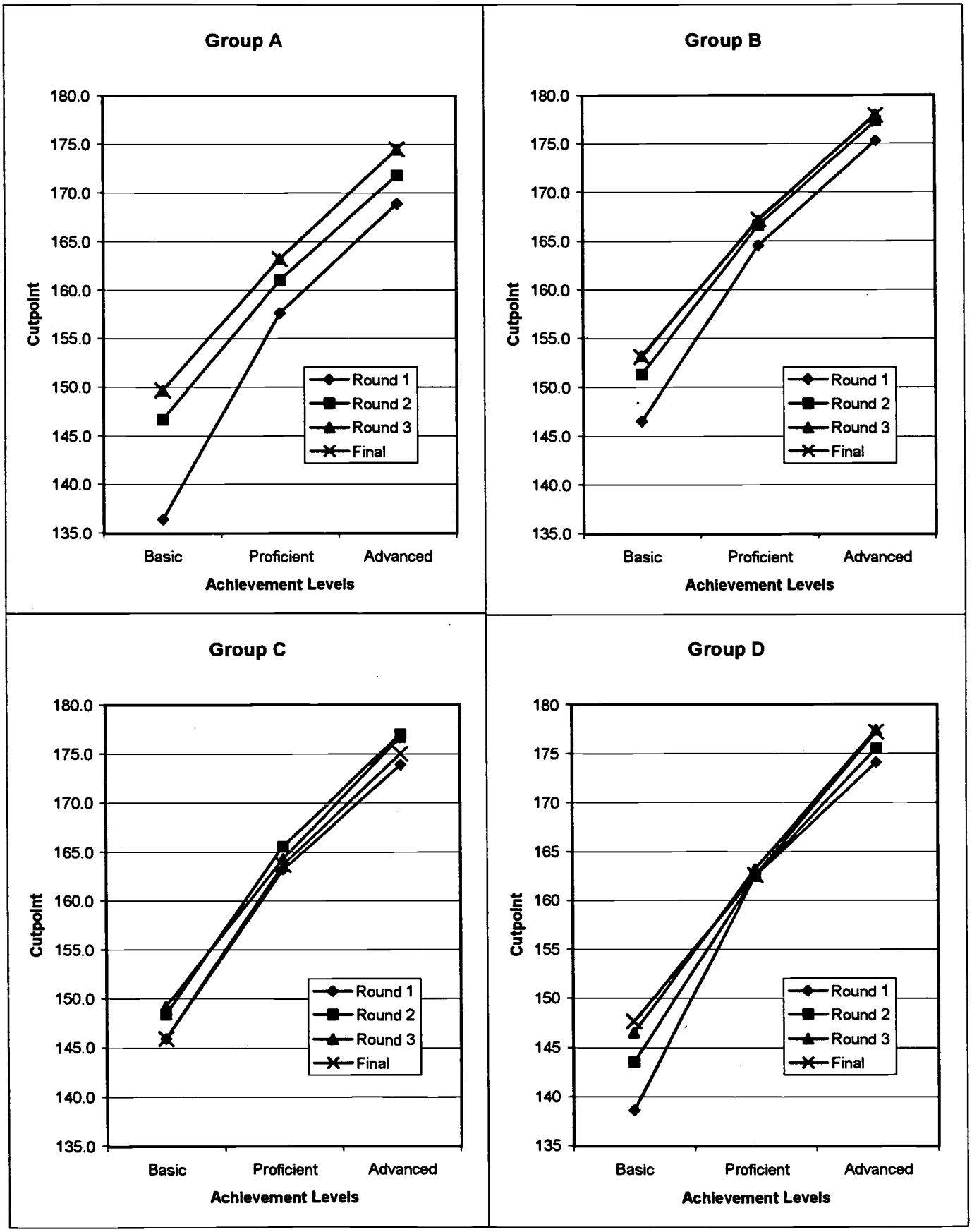
Mean Proficient Cutpoints for Rating Method by Type of Feedback



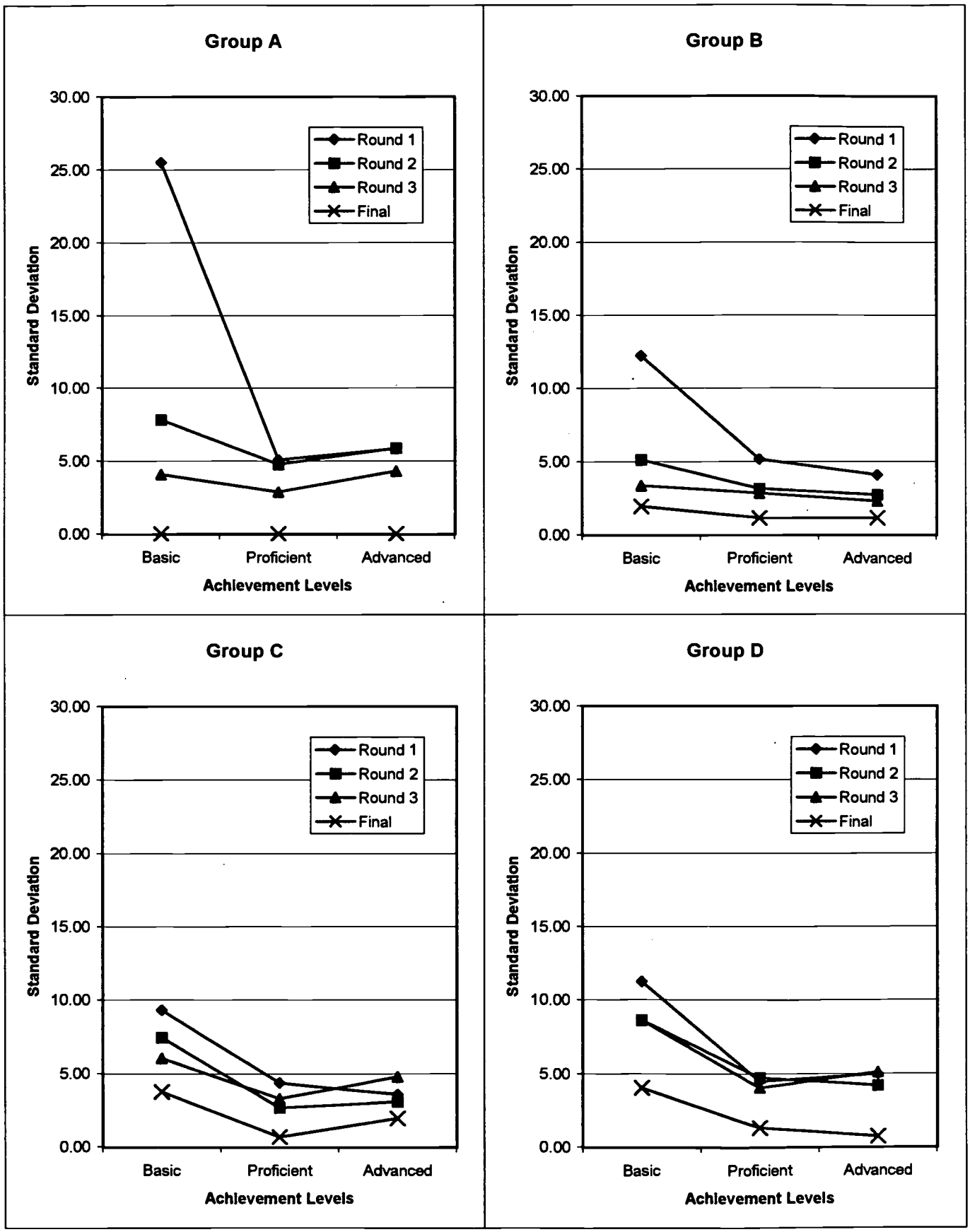
Mean Advanced Cutpoints for Rating Method by Type of Feedback



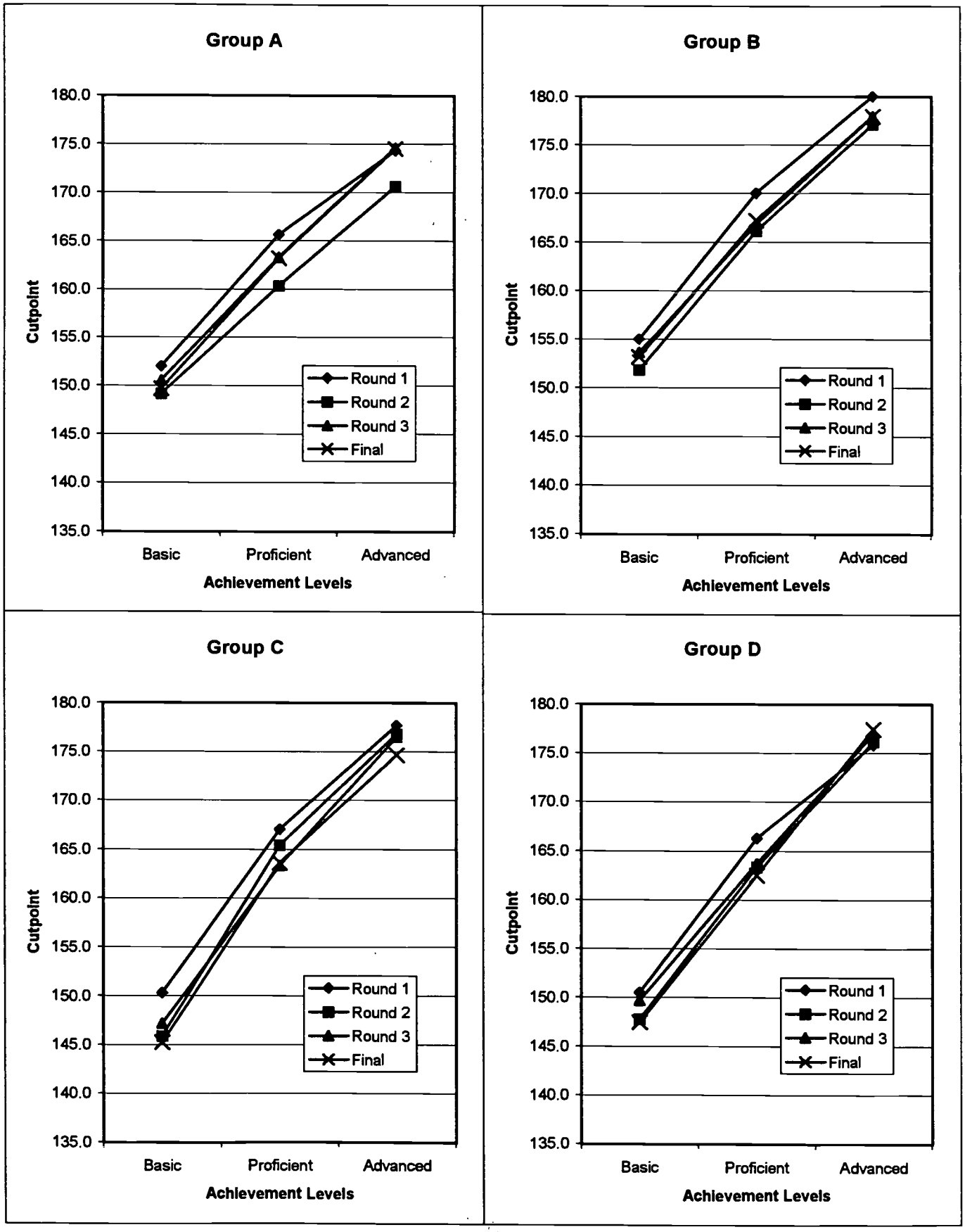
Civics Field Trial 2
Cutpoints Set by Different Groups on Different Rounds



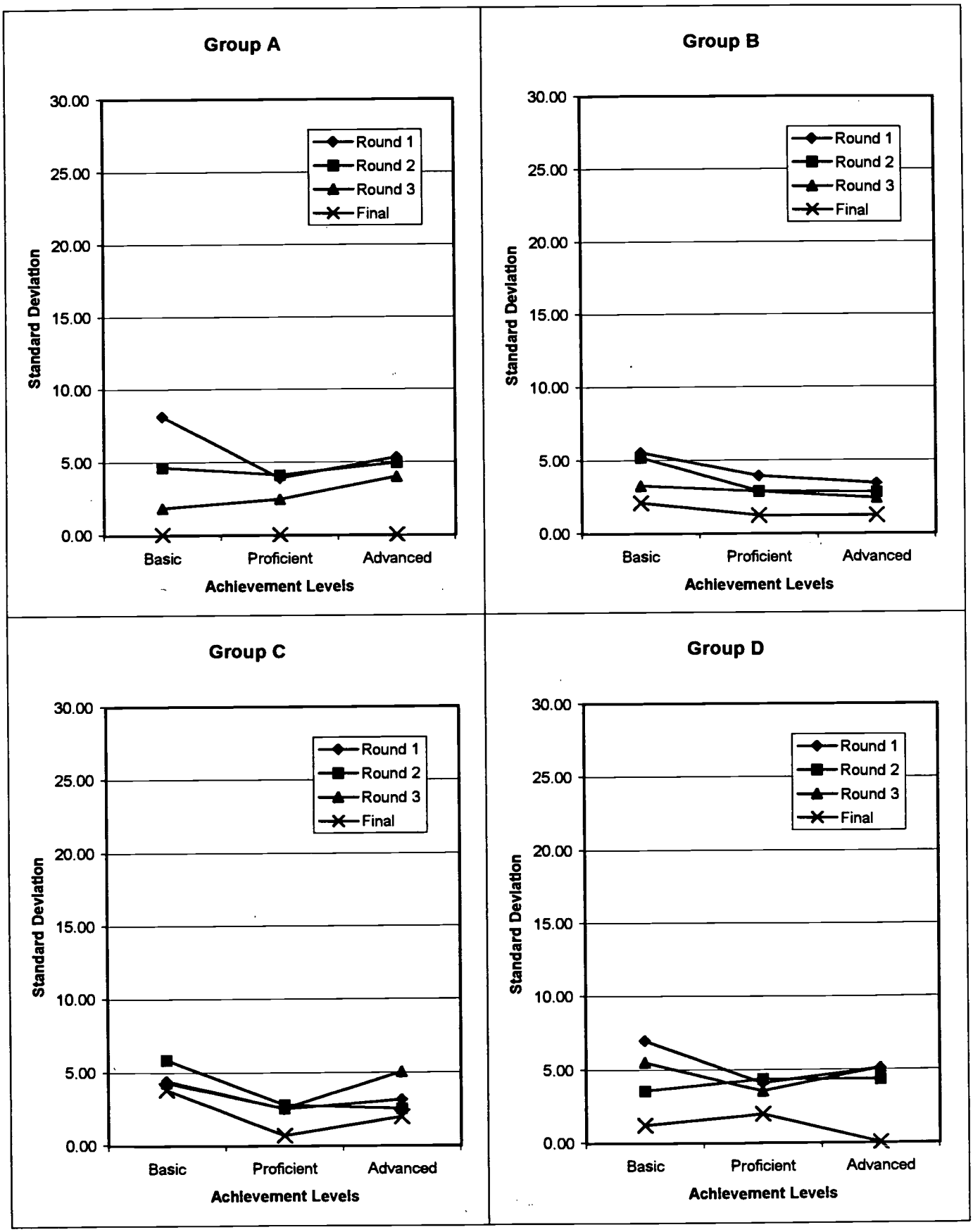
Civics Field Trial 2
Standard Deviations of Cutpoints Set by Different Groups on Different Rounds



Civics Field Trial 2
Cutpoints Set by Different Groups on Different Rounds
Without the Possible Outliers



Civics Field Trial 2
Standard Deviations of Cutpoints Set by Different Groups on Different Rounds
Without the Possible Outliers



Writing Field Trial 2
Descriptive Statistics—Outliers Replaced with Means

Means and Standard Deviations Across All Groups and Rounds

Variable	N	Mean	SD
Basic	40	129.8	13.1
Proficient	40	161.5	14.2
Advanced	40	200.8	14.3

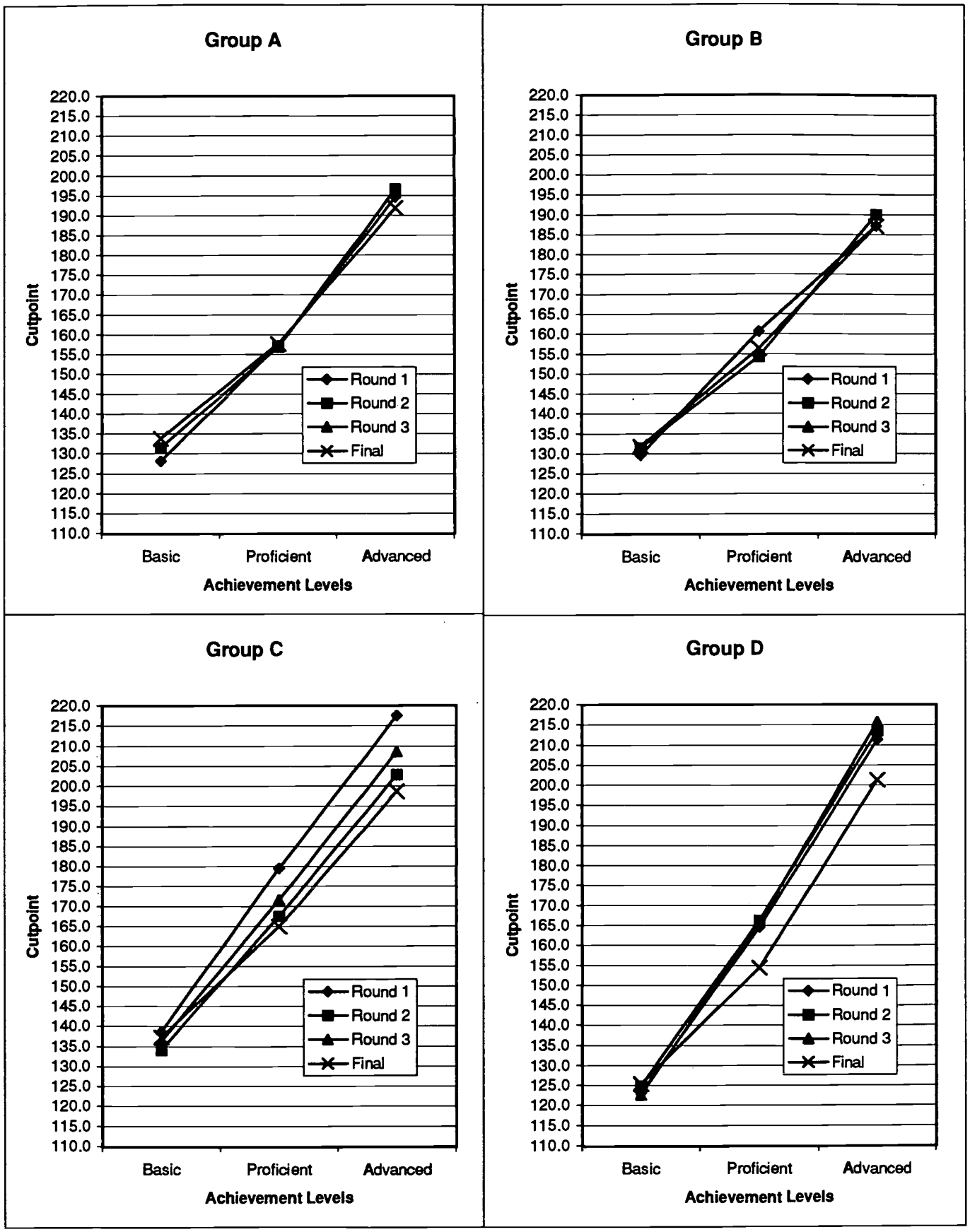
Means and Standard Deviations by Rounds Across All Groups

Round	Variable	N	Mean	SD
1	Basic	40	129.4	13.2
	Proficient	40	165.5	12.8
	Advanced	40	203.8	18.7
2	Basic	40	129.8	13.1
	Proficient	40	161.5	14.2
	Advanced	40	200.8	14.3

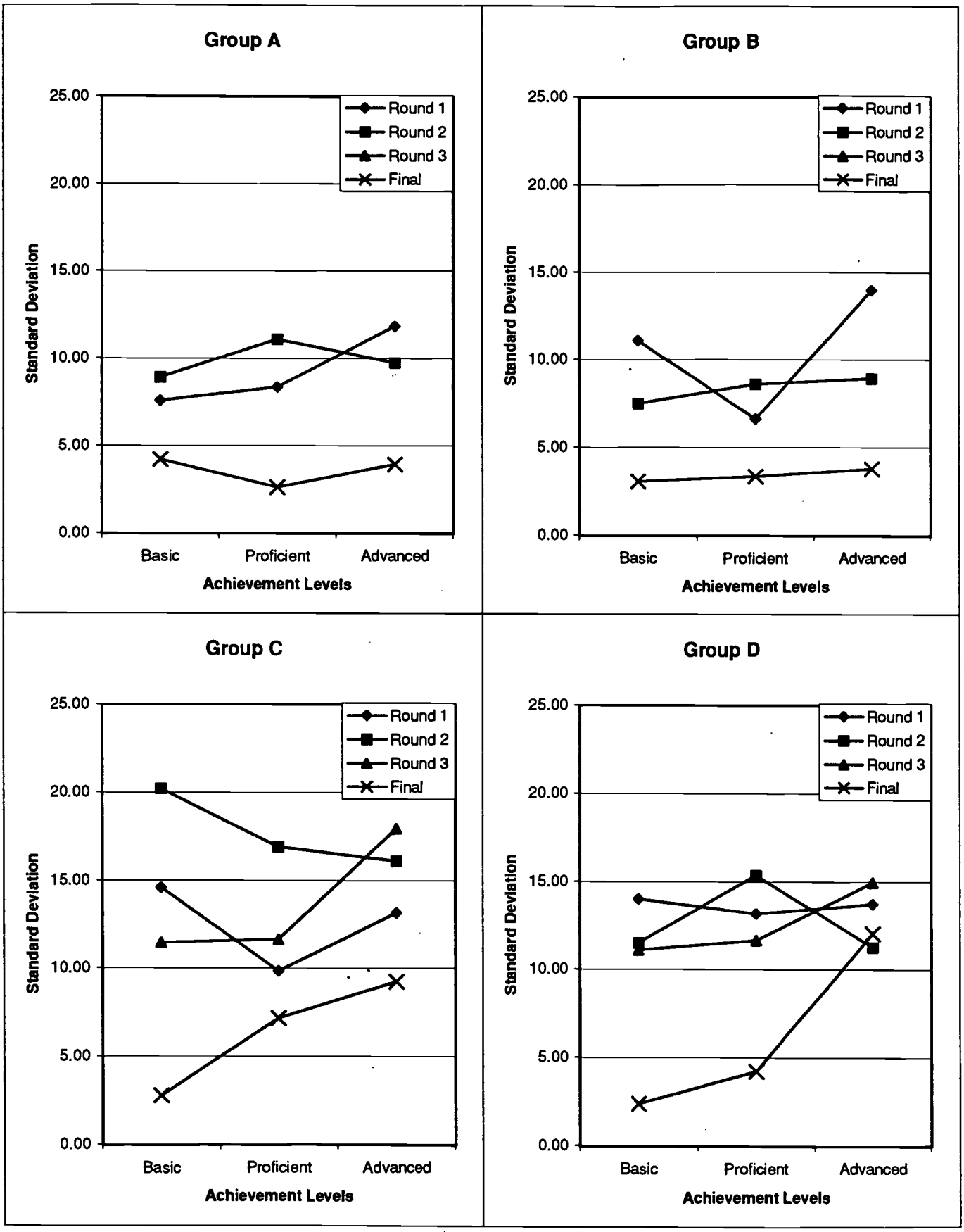
Means and Standard Deviations by Groups Across All Rounds

Group	Variable	N	Mean	SD
A	Basic	20	129.8	7.9
	Proficient	20	157.0	9.5
	Advanced	20	195.5	10.6
B	Basic	20	130.6	9.3
	Proficient	20	157.5	8.1
	Advanced	20	188.6	11.5
C	Basic	20	135.4	17.8
	Proficient	20	173.8	14.7
	Advanced	20	212.4	16.9
D	Basic	20	122.5	12.5
	Proficient	20	165.9	13.9
	Advanced	20	212.8	12.2

Field Trial 2 for Writing
Cutpoints Set by Different Groups on Different Rounds



Field Trial 2 for Writing
Standard Deviations of Cutpoints Set
by Different Groups on Different Rounds



1998 NAEP Achievement Levels-Setting Process
Field Trial 2 for Civics
Summary of Responses to Process Evaluation Questions

Means and Frequencies of Responses to Questions Related to Ratings Using Different Methods

Questions	Round	A (n=10) ME/IM w/ CD						B (n=11) ME/IM w/o CD						C (n=11) ME/MR w/ CD						D (n=11) ME/MR w/o CD					
		5	4	3	2	1	Mean	5	4	3	2	1	Mean	5	4	3	2	1	Mean	5	4	3	2	1	Mean
1. The instructions on what I was to do during the first/second rating session were: (5=Absolutely Clear; 1=Not at all Clear)	1	1	6	0	3	0	3.50	2	5	3	1	0	3.73	3	4	3	1	0	3.81	2	5	3	1	0	3.73
	2	7	3	0	0	0	4.70	7	4	0	0	0	4.63	5	5	1	0	0	4.36	7	4	0	0	0	4.64
	3	0	7	2	1	0	3.60	5	4	2	0	0	4.27	7	3	1	0	0	4.55	7	3	1	0	0	4.54
2. My level of understanding of the tasks I was to accomplish during the first/second rating session was: (5=Totally Adequate; 1=Totally Inadequate)	1	1	4	4	1	0	3.50	1	8	1	1	0	3.82	5	3	2	0	1	4.00	2	6	3	0	0	3.91
	2	5	5	0	0	0	4.50	6	5	0	0	0	4.55	4	5	2	0	0	4.18	7	4	0	0	0	4.64
	3	0	7	3	0	0	3.70	6	4	0	1	0	4.36	6	4	1	0	0	4.45	6	4	1	0	0	4.45
3. The amount of time I had to complete the tasks I was to accomplish during the first/second rating session was: (5=Far too Long; 1=Far too Short)	1	0	2	7	1	0	3.10	2	3	5	1	0	3.55	0	3	7	1	0	3.18	0	0	10	1	0	2.91
	2	0	5	5	0	0	3.50	3	5	3	0	0	4.00	0	1	6	3	1	2.64	0	0	7	2	2	2.45
	3	0	3	7	0	0	3.30	1	1	9	0	0	3.27	1	0	8	1	1	2.91	0	0	10	1	0	2.91
4. The most accurate description of my level of confidence in the ratings I provided to represent the three achievement levels during the first/second rating session is that I was: (5=Totally Confident; 1=Not at all Confident)	1	0	3	6	1	0	3.20	0	4	4	3	0	3.09	0	5	4	1	1	3.18	1	4	4	2	0	3.36
	2	1	8	1	0	1	4.00	2	5	3	0	1	3.64	1	8	1	1	0	3.82	2	6	1	0	2	3.55
	3	0	6	4	0	0	3.60	1	9	0	1	0	3.91	0	8	1	2	0	3.55	2	6	1	1	1	3.64
5. The method for rating multiple-choice items was conceptually clear. (5=Totally Agree; 1=Totally Disagree)	1	3	5	2	0	0	4.10	4	5	2	0	0	4.18	5	3	3	0	0	4.18	7	4	0	0	0	4.63
	2	6	4	0	0	0	4.60	7	4	0	0	0	4.64	2	7	2	0	0	4.00	4	7	0	0	0	4.36
	3	2	2	5	1	0	3.50	5	5	1	0	0	4.36	4	5	2	0	0	4.18	2	7	2	0	0	4.00
6. The method for rating multiple-choice items was easy to apply. (5=Totally Agree; 1=Totally Disagree)	1	3	4	3	0	0	4.00	3	3	4	1	0	3.73	4	3	4	0	0	4.00	2	6	3	0	0	3.91
	2	5	5	0	0	0	4.50	8	0	2	1	0	4.36	2	6	3	0	0	3.91	3	8	0	0	0	4.27
	3	2	2	5	1	0	3.50	5	4	2	0	0	4.27	4	3	4	0	0	4.00	3	5	3	0	0	4.00
7. The method for rating constructed-response items was conceptually clear. (5=Totally Agree; 1=Totally Disagree)	1	0	5	3	1	0	3.20	2	7	1	1	0	3.91	1	5	5	0	0	3.64	5	4	1	1	0	4.18
	2	5	4	1	0	0	4.40	4	5	1	1	0	4.09	1	7	3	0	0	3.82	3	5	3	0	0	4.00
	3	2	2	5	1	0	3.50	4	5	2	0	0	4.18	3	4	4	0	0	3.91	1	7	2	1	0	3.73
8. The method for rating constructed-response items was easy to apply. (5=Totally Agree; 1=Totally Disagree)	1	0	4	2	3	1	2.90	2	4	4	1	0	3.64	2	1	7	1	0	3.36	1	5	4	1	0	3.55
	2	3	5	2	0	0	4.10	3	4	3	0	1	3.73	1	7	3	0	0	3.81	2	5	4	0	0	3.81
	3	2	3	4	1	0	3.60	4	5	2	0	0	4.18	3	4	4	0	0	3.91	2	5	3	1	0	3.73

Means and Frequencies of Responses to Questions Related to the Item Mapping Method

Questions	A (n=10) ME/IM w/ CD						B (n=11) ME/IM w/o CD					
	5	4	3	2	1	Mean	5	4	3	2	1	Mean
1. Information about the item map for each item to be considered during the third rating session was: (5=Absolutely Clear; 1=Not at All Clear)	1	5	4	0	0	3.70	5	4	2	0	0	4.27
2. Instructions on the interpretation of the item map for each item to be considered during the third rating session were: (5=Absolutely Clear; 1=Not at All Clear)	1	3	6	0	0	3.50	6	4	1	0	0	4.45
3. Instructions on the used of the item for each item to be considered during the third rating session was: (5=Absolutely Clear; 1=Not at All Clear)	1	5	4	0	0	3.70	6	3	2	0	0	4.36
4. The relationship between my item-by-item ratings and the item map was: (5=Absolutely Clear; 1=Not at All Clear)	1	6	3	0	0	3.80	6	4	1	0	0	4.45
5. If the length of time spent on instructions concerning the item map for each item were to be changed, I would recommend: (5=Far More Time; 1= Far Less Time)	0	1	8	1	0	3.00	0	1	9	0	0	3.10
6. Following the instructions, the most accurate description of my level of confidence in my ability to use the item map for each item during the third rating session was: (5= Totally Confident; 1=Not at All Confident)	0	6	4	0	0	3.60	2	8	1	0	0	4.09
7. The cutscores I set using item maps are likely to be much closer to my concept of Borderline Basic performance than my previous ratings. (5=Totally Agree; 1=Totally Disagree)	2	5	3	0	0	3.90	7	3	1	0	0	4.54
8. The cutscores I set using item maps are likely to be much closer to my concept of Borderline Proficient performance than my previous ratings. (5=Totally Agree; 1=Totally Disagree)	2	5	3	0	0	3.90	8	2	1	0	0	4.64
9. The cutscores I set using item maps are likely to be much closer to my concept of Borderline Advanced performance than my previous ratings. (5=Totally Agree; 1=Totally Disagree)	2	5	3	0	0	3.90	8	2	1	0	0	4.64

Means and Frequencies of Responses to Questions Related to the Reckase Method (Round 2)

Questions	C (n=11) ME/MR w/ CD						D (n=11) ME/MR w/o CD					
	5	4	3	2	1	Mean	5	4	3	2	1	Mean
1. Information about the source of the data in the MR charts for each item to be considered during the second rating session was: (5=Absolutely Clear; 1= No at All Clear)	6	3	2	0	0	4.36	2	7	2	0	0	4.00
2. Instructions on the interpretation of data in the MR charts for each item to be considered during the second rating session were: (5=Absolutely Clear; 1= No at All Clear)	6	4	1	0	0	4.54	6	5	0	0	0	4.55
3. Instructions on the used of data in the MR charts for each item to be considered during the second rating session were: (5=Absolutely Clear; 1= No at All Clear)	6	3	2	0	0	4.36	5	6	0	0	0	4.45
4. The relationship between my item-by-item ratings and the MR charts was: (5=Absolutely Clear; 1= No at All Clear)	5	4	1	1	0	4.18	4	6	1	0	0	4.27
5. If the length of time on instructions concerning the MR charts for each item were to be changed, I would recommend: (5=Far More Time; 1=Far Less Time)	1	1	8	1	0	3.18	4	3	2	0	0	3.00
6. Following instruction, the most accurate description of my level of confidence in my ability to use the MR charts for the second/third rating session is that I was: (5= Totally Confident; 1= Not at All Confident)	3	4	4	0	0	3.91	3	6	2	0	0	4.09
7. Marking the MR charts revealed patterns in my ratings for multiple-choice and constructed-response items. (5=Totally Agree; 1=Totaly Disagree)	3	4	1	2	1	3.55	5	4	1	0	1	4.09
8. Marking the MR charts revealed patterns in my ratings based on item content. (5=Totally Agree; 1=Totaly Disagree)	1	5	1	3	1	3.18	3	4	2	2	0	3.73
9. Marking the MR charts revealed patterns of consistency/inconsistency in my ratings. (5=Totally Agree; 1=Totaly Disagree)	1	6	2	1	1	4.45	4	5	2	0	0	4.18
10. The cutscores computed from my round 2 ratings using the MR charts are likely to be much closer to my concept of borderline Basic performance than my previous ratings. (5=Totally Agree; 1=Totaly Disagree)	1	5	4	1	0	3.55	4	5	1	0	1	4.00
11. The cutscores computed from my round 2 ratings using the MR charts are likely to be much closer to my concept of borderline Proficient performance than my previous ratings. (5=Totally Agree; 1=Totaly Disagree)	2	3	5	1	0	3.55	5	2	3	0	1	3.91
12. The cutscores computed from my round 2 ratings using the MR charts are likely to be much closer to my concept of borderline Advanced performance than my previous ratings. (5=Totally Agree; 1=Totaly Disagree)	1	4	5	1	0	3.45	6	1	3	0	1	4.00

Means and Frequencies of Responses to Questions Related to the Reckase Method (Round 3)

Questions	C (n=11) ME/MR w/ CD						D (n=11) ME/MR w/o CD					
	5	4	3	2	1	Mean	5	4	3	2	1	Mean
1. I was confident in my selection of one row on the MR charts to represent my ratings for Basic. (5=Totally Agree; 1=Totally Disagree)	1	5	4	1	0	3.55	1	6	3	1	0	3.64
2. I was confident in my selection of one row on the MR charts to represent my ratings for Proficient. (5=Totally Agree; 1=Totally Disagree)	2	6	2	1	0	3.82	3	5	2	1	0	3.91
3. I was confident in my selection of one row on the MR charts to represent my ratings for Advanced. (5=Totally Agree; 1=Totally Disagree)	2	6	2	1	0	3.82	1	7	2	1	0	3.73
4. I would have preferred selection a range of rows instead of a single row on the MR charts to represent my ratings for Basic. (5=Totally Agree; 1=Totally Disagree)	3	3	3	1	1	3.55	3	3	2	2	1	3.45
5. I would have preferred selection a range of rows instead of a single row on the MR charts to represent my ratings for Proficient. (5=Totally Agree; 1=Totally Disagree)	2	3	3	2	1	3.27	3	1	2	4	1	3.09
6. I would have preferred selection a range of rows instead of a single row on the MR charts to represent my ratings for Advanced. (5=Totally Agree; 1=Totally Disagree)	2	3	3	2	1	3.27	3	1	3	3	1	3.18
7. I would have preferred averaging the scores of the rows instead of selecting a single row on the MR charts to represent my ratings. (5=Totally Agree; 1=Totally Disagree)	2	3	3	2	1	3.27	3	0	4	2	2	3.00
8. The cutscores I set using the MR charts are much closer to my concept of Borderline Basic performance than my previous ratings. (5=Totally Agree; 1=Totally Disagree)	0	5	4	2	0	3.27	2	2	5	2	0	3.36
9. The cutscores I set using the MR charts are much closer to my concept of Borderline Proficient performance than my previous ratings. (5=Totally Agree; 1=Totally Disagree)	0	7	2	2	0	3.45	1	2	6	2	0	3.18
10. The cutscores I set using the MR charts are much closer to my concept of Borderline Advanced performance than my previous ratings. (5=Totally Agree; 1=Totally Disagree)	0	8	3	0	0	3.73	0	4	5	1	1	3.09

Means and Frequencies of Responses to Questions Related Feedback

Questions	Round	A (n=10) ME/IM w/ CD						B (n=11) ME/IM w/o CD						C (n=11) ME/MR w/ CD						D (n=11) ME/MR w/o CD					
		5	4	3	2	1	Mean	5	4	3	2	1	Mean	5	4	3	2	1	Mean	5	4	3	2	1	Mean
1. When I provided ratings during the second/third rating session, my judgments were influenced by student performance data reporting the % correct or average on each item: (5=Greatly; 1=Not at All)	2	4	2	4	0	0	4.00	5	4	2	0	0	4.27	0	1	7	2	1	2.73	2	4	5	0	0	3.73
	3	0	1	5	3	1	2.60	3	1	5	2	0	3.45	0	2	2	3	4	2.18	0	3	2	4	2	2.55
2. When I provided ratings during the second/third rating session, my judgments were influenced by rater location data: (5=Greatly; 1=Not at All)	2	3	3	2	2	0	4.00	2	5	4	0	0	3.82	1	3	5	2	0	3.27	1	3	6	1	0	3.36
	3	1	3	5	1	0	3.40	2	3	2	4	0	3.27	0	4	4	3	0	3.09	3	5	2	1	0	3.91
3. When I provided ratings during the second/third rating session, my judgments were influenced by the Whole Booklet Exercise and Feedback: (5=Greatly; 1=Not at All)	2	0	2	5	1	2	3.70	1	4	5	1	0	3.45	0	2	3	6	0	2.64	0	3	4	2	2	2.73
	3	0	2	5	2	1	2.80	2	2	3	3	1	3.09	0	4	1	4	2	2.64	0	2	4	3	2	2.54
4. When I provided ratings during the second/third rating session, my judgments were influenced by the [data in the MR charts/item maps]: (5=Greatly; 1=Not at All)	2													3	5	3	0	0	4.00	1	7	3	0	0	3.82
	3	0	5	4	1	0	3.40	3	7	0	0	0	4.30	4	4	1	2	0	3.91	5	3	2	1	0	4.09
5. The most useful information was the P-Value data, i.e., % of students who correctly answered or the average score on each item. (5=Totally Agree; 1=Totally Disagree)	2	3	3	4	0	0	2.70	4	6	1	0	0	4.27	0	4	5	1	1	3.09	3	2	5	1	0	3.64
	3	0	1	6	2	1	2.70	0	5	4	2	0	3.27	0	0	6	2	3	2.27	0	3	2	6	0	2.73
6. The most useful information was the report on the location of my cutpoint at each achievement level relative to others in my group (rater location data). (5=Totally Agree; 1=Totally Disagree)	2	0	4	5	1	0	3.30	1	3	7	0	0	3.45	1	6	4	0	0	3.73	2	2	3	3	1	3.09
	3	0	6	4	0	0	3.60	1	5	2	3	0	3.36	0	7	4	0	0	3.64	0	4	4	2	1	3.00
7. The most useful information was the report on student performance via the Whole Booklet Exercise and Feedback. (5=Totally Agree; 1=Totally Disagree)	2	0	2	3	3	2	3.30	0	3	5	3	0	3.00	0	0	6	5	0	2.55	0	2	7	2	0	3.00
	3	0	1	3	5	1	2.40	0	3	4	4	0	2.91	0	2	5	3	1	2.73	0	2	4	3	2	2.55
8. The most useful information was the [MR charts/item maps]. (5=Totally Agree; 1=Totally Disagree)	2													0	7	3	1	0	3.55	2	4	4	1	0	3.64
	3	0	5	3	2	0	3.30	2	9	0	0	0	4.18	3	5	1	2	0	3.82	6	4	1	0	0	4.45
9. I used the P-Value data to adjust my ratings during Round 2/3. (5=To a Great Extent; 1=Not at All)	2	5	2	3	0	0	2.50	5	6	0	0	0	4.45	0	2	7	1	1	2.90	3	2	5	1	0	3.64
	3	0	3	3	1	3	2.60	0	2	4	2	3	2.45	0	0	3	3	5	1.82	0	1	3	6	1	2.36
10. I used the rater location data to adjust my ratings during Round 2/3. (5=To a Great Extent; 1=Not at All)	2	1	3	4	2	0	4.20	1	3	6	1	0	3.36	1	6	4	1	1	3.73	1	2	7	1	0	3.27
	3	1	4	4	0	1	3.40	0	3	5	3	0	3.00	0	7	2	1	1	3.36	1	6	2	2	0	3.55
11. I used the Whole Booklet data to adjust my ratings during Round 2/3. (5=To a Great Extent; 1=Not at All)	2	0	1	3	3	3	3.30	1	3	3	4	0	3.09	0	0	6	6	1	2.27	0	3	5	2	1	2.91
	3	0	1	5	2	2	2.50	1	1	3	3	3	2.45	0	2	3	3	2	2.55	0	0	5	4	2	2.27

Averages of Responses to Questions Related to the ALS Process

Questions		A (n=10) ME/IM w/ CD						B (n=11) ME/IM w/o CD						C (n=11) ME/MR w/ CD						D (n=11) ME/MR w/o CD					
		5	4	3	2	1	Mean	5	4	3	2	1	Mean	5	4	3	2	1	Mean	5	4	3	2	1	Mean
1) The most accurate description of my level of confidence in the achievement levels ratings I provided was: (5=Totally Confident; 1=Not at all Confident)		2	5	3	0	0	3.90	3	6	2	0	0	4.09	1	9	0	1	0	3.91	2	8	1	0	0	4.09
2) I would describe the effectiveness of this achievement levels-setting process as: (5=Highly Effective; 1=Not at all Effective)		0	7	3	0	0	3.70	4	6	0	1	0	4.18	1	3	5	2	0	3.27	0	8	3	0	0	3.73
3) I feel that this NAEP ALS process provided me an opportunity to use my best judgment in rating items to set achievement levels for the NAEP Geography Assessment: (5=To a Great Extent; 1=Not at All)		1	4	5	0	0	3.60	5	5	1	0	0	4.36	1	5	4	1	0	3.55	4	5	2	0	0	4.18
4) I feel that this NAEP ALS process produced achievement levels that are defensible: (5=To a Great Extent; 1=Not at All)		1	6	2	1	0	3.70	4	6	0	1	0	4.18	3	5	2	1	0	3.91	3	6	1	1	0	4.00
5) I feel that this NAEP ALS process produced achievement levels that will generally be considered reasonable: (5=To a Great Extent; 1=Not at All)		3	6	0	1	0	4.10	8	2	0	1	0	4.55	5	2	4	0	0	4.09	4	4	2	1	0	4.00
6) I would be willing to sign a statement (after reading it, of course) recommending the use of achievement levels resulting from this ALS procedure:	Yes, definitely.	1					10%	7					64%	3					27%	3					27%
	Yes, probably.	8					80	4					36	7					64	8					73
	No, probably not.	1					10	0					0	1					9	0					0
	No, definitely not.	0					0	0					0	0					0	0					0

Sample Reckase Chart

	1	2	3	4	5	6	7	8	9	10	11	12
AA	99	99	99	3.0	99	99	99	99	99	99	99	3.0
AB	99	99	99	3.0	99	99	99	99	99	99	99	3.0
AC	99	99	99	3.0	99	99	99	99	99	99	99	3.0
AD	99	99	99	*3.0*	99	99	99	99	99	99	99	3.0
AE	99	99	99	3.0	99	99	99	99	99	99	99	3.0
AF	99	99	99	3.0	99	99	99	99	99	99	99	3.0
AG	99	*99*	99	3.0	99	99	99	99	99	99	99	3.0
AH	99	*99*	*99*	2.9	99	99	99	*99*	99	99	99	*3.0*
AI	99	99	*99*	2.9	99	99	99	99	99	99	99	*3.0*
AJ	99	99	99	*2.9*	99	99	99	99	*99*	99	*99*	3.0
AK	99	99	99	*2.9*	99	*99*	99	99	*99*	99	*99*	3.0
AL	99	99	99	2.9	*99*	*99*	*99*	99	99	*99*	99	3.0
AM	99	99	99	2.9	99	99	99	99	99	99	99	3.0
AN	99	99	99	2.8	99	99	99	99	99	99	99	3.0
AO	*99*	98	99	2.8	99	99	99	99	99	99	99	3.0
AP	*99*	98	99	2.8	99	99	99	98	99	99	99	3.0
AQ	99	*98*	98	2.7	99	99	99	98	99	99	99	2.9
AR	99	98	98	2.7	99	99	99	98	99	99	99	2.9
AS	99	98	*98*	2.6	99	99	99	98	99	99	99	*2.9*
AT	99	97	98	2.6	99	99	99	97	99	99	99	2.9
AU	99	97	98	2.5	99	99	99	97	98	99	98	2.9
AV	99	97	97	2.4	99	99	99	97	98	99	98	2.8
AW	99	96	97	2.4	99	98	99	96	98	99	98	2.8
AX	99	96	97	2.3	98	98	98	96	97	98	97	2.7
AY	99	96	96	2.2	98	97	98	95	96	97	96	2.6
AZ	99	95	96	2.1	97	95	98	94	96	95	96	2.5
BA	99	94	95	2.0	96	94	98	93	95	91	95	2.4
BB	99	94	94	1.9	94	91	97	92	94	85	93	2.2
BC	99	93	94	1.9	92	87	97	91	92	76	91	2.1
BD	99	92	93	1.8	88	82	96	90	91	65	89	1.9
BE	98	91	92	1.7	84	76	96	88	89	53	87	1.8
BF	98	90	91	1.6	79	69	95	86	86	43	84	1.6
BG	98	89	89	1.5	72	61	94	84	84	36	80	1.5
BH	97	88	88	1.5	66	53	93	82	81	31	76	1.4
BI	97	87	87	1.4	59	46	92	80	77	28	71	1.3
BJ	96	85	85	1.4	54	40	91	77	73	27	66	1.2
BK	95	84	83	1.3	49	35	89	75	69	26	61	1.2
BL	94	82	81	1.3	46	31	87	72	64	25	56	1.1
BM	93	80	79	1.2	43	28	85	69	60	25	51	1.1
BN	91	78	77	1.2	41	26	83	65	55	25	46	*1.1*
BO	89	76	74	1.2	40	25	81	62	51	25	42	1.1
BP	86	74	72	1.1	39	24	78	59	46	25	38	1.1
BQ	83	72	69	1.1	38	23	75	55	42	25	35	1.0
BR	80	69	66	*1.1*	38	23	72	52	38	25	32	1.0
BS	76	67	63	*1.1*	38	22	68	49	35	25	29	1.0
BT	72	64	60	1.1	37	22	65	45	32	25	27	1.0
BU	68	62	57	1.1	37	22	61	42	30	25	26	1.0
BV	63	59	54	1.0	37	22	57	40	28	25	25	1.0
BW	58	56	51	1.0	37	22	53	37	26	25	24	1.0
BX	53	54	48	1.0	37	22	50	34	24	25	23	1.0
BY	48	51	45	1.0	37	22	46	32	23	25	22	1.0
BZ	44	49	43	1.0	37	22	43	30	22	25	22	1.0
CA	40	46	40	1.0	37	22	40	28	21	25	22	1.0
CB	37	44	38	1.0	37	22	37	27	21	25	21	1.0
CC	34	42	36	1.0	37	22	34	25	20	25	21	1.0
CD	31	40	34	1.0	37	22	32	24	20	25	21	1.0

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM029935

REPRODUCTION RELEASE

(Specific Document)

NCME

I. DOCUMENT IDENTIFICATION:

Title: FIELD TRIALS TO DETERMINE WHICH RATING METHOD(S) TO USE IN THE 1998 NAEP ACHIEVEMENT LEVELS-SETTING PROCESS FOR CIVICS AND WRITING

Author(s): SUSAN COOPER LOOMIS, LUZ BAY, WEN-LING YANG, AND PATRICIA HANICK

Corporate Source: ACT, INC.

Publication Date: 4/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

[X]

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

[ ]

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

[ ]

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Susan C. Loomis

Printed Name/Position/Title: SUSAN COOPER LOOMIS, DIRECTOR, NAEP PROJEC

Organization/Address: ACT, INC., 2255 NORTH DUBUQUE ROAD, P.O. BOX 168, IOWA CITY, IA 52243-0168

Telephone: 319/337-1048 FAX: 319/337-1497

E-Mail Address: LOOMLS@ACT.ORG Date: 5/27/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com